단백질 상호작용 네트워크을 위한 개념 기반 추상화

최재훈[O+] 박종민[+] 김기헌[*] 박선희[+]
한국전자통신연구원
{jhchoi[O+], jmpark93[+], psh[+]}@etri.re.kr; khkim[*]@chonbuk.ac.kr

A Concept-Based Approach for Abstracting Protein Interaction Networks

Jae-Hun Choi[O+] , Jong-Min Park[+], Ki-Heon Kim[*], Seon-Hee Park[+]
[+] Electronic Telecommunication Research Institute(ETRI)
[*] Chonbuk National University

요 약

본 논문은 세포 내에 존재하는 방대한 단백질들 사이의 상호작용 관계 네트워크에서 생물학적인 의미 연관성을 가지는 부분 네트워크를 콤포지트로 추상화할 수 있는 방법을 제안한다. 이 추상화를 위해 네트워크에서 구조적으로 완전한 부분 네트워크, 개념적으로 인접한 부분 네트워크 그리고 두 조건을 모두 만족하는 부분 네트워크를 탐색한다. 따라서, 사용자는 방대한 네트워크을 개념적인 관점에서 분석할 수 있으며, 특정한 의미을 가지는 부분 네트워크를 쉽게 검색할 수 있다.

1. 서론

현대 생명과학 연구는 개개의 유전자나 단백질보다 이들 사이의 복잡한 상호작용을 통해 전체적인 관점에서 생명 현상을 규명하려는데 초점을 맞추고 있다. 단백질은 DNA에 포함된 유전자가 발현되어 생성되는 물질로서, 다른 단백질과의 유기적인 상호작용을 통해 다양한 생명현상에 주도적 역할을 하게 된다. 즉, 신호전달, 세포 생명 주기, 세포 발달, DNA 복제, 물질대사 등에 핵심적으로 관여하여 세포의 생리활성 반응을 조절하게 된다. 이 상호작용을 단백질들 사이의 관계로 나타내면 하나의 네트워크 형태로 표현될 수 있다.

단백질 상호작용 네트워크는 단백질들 사이의 관계들에 대한 집합으로 정의할 수 있다. 즉, 단백질은 네트워크의 노드 그리고 단백질들 사이의 상호작용은 노드들 사이의 관계로 각각 표현된다. 따라서, 두 단백질들 사이의 관계는 하나의 단백질이 다른 단백질과 특정한 생물학적 작용을 한다고 해석할 수 있다.

현재, 생물 종에 따라 많은 단백질 상호작용 네트워크가 이스트 투 하이브리드(yeast two hybrid)와 같은 고성능 생물 실험(high-throughput screening)을 통해 추출되고 있다[1]. 대표적으로 이와 같은 실험을 통해 구축된 데이터로는 PIM, BIND, DIP, GRID 등이 있다.

일반적으로 네트워크는 매우 방대한 단백질들과 이들 사이의 복잡한 관계로 표현된다. 예를 들어, DIP에서 밝혀낸 Saccharomyces Cerevisiae의 네트워크는 4,772 단백질 그리고 15,479 상호작용 관계로 구성되어 있다. 따라서, 하나의 네트워크에 포함된 방대한 단백질들을 생물학자가 개별적으로 분석하기란 매우 어렵다.

Ospreyl나 Cytoscape와 같은 기존 시스템은 특정 단백질과 상호관계를 가지는 다른 단백질을 파악할 수는 있도록 지원한다. 그러나, 단백질 콤플렉스와 같이 유사한 기능을 하는 주위 단백질들을 하나의 단위로하는 상호관계를 파악할 수 없다. 하나의 단백질은 그 자체로서 기능을 하지만, 주변의 다른 단백질과의 밀접한 상호작용을 통해 특정한 생물학적인 역할을 수행한다. 이 역할을 수행하는 단백질들을 본 논문에서는 바이오 콤포지트로 정의한다.

단백질 콤포지트를 탐색하기 위한 연구들은 일반적으로 네트워크 토플로지 정보를 이용한다. 대표적인 방법으로 네트워크에 포함된 클릭 구조의 부분 네트워크를 탐색하여 콤포지트를 밝혀내고 있다. 클릭 구조 네트워크에서 하나의 노드는 다른 모든 노드들과 상호관계를 가지게 되기 때문에 클릭 구조의 콤포지트에 포함된 단백질들은 서로 매우 밀접한 의미적 관계를 가지고 있다.

그러나, 일반적인 콤포지트는 반드시 클릭 구조를 가지고 있지 않으며, 단지 단백질들 사이에 매우 밀접한 의미적 관계만를 가지고 있다. 따라서, 클릭의 의한 콤포지트 탐색 방법은 의미적인 연관성이 없는 많은 클릭을 탐색할 수 있는 단점(false positive) 그리고, 클릭 구조가 아닌 매우 많은 다른 콤포지트는 탐색할 수 없는 단점(false negative)을 가진다.

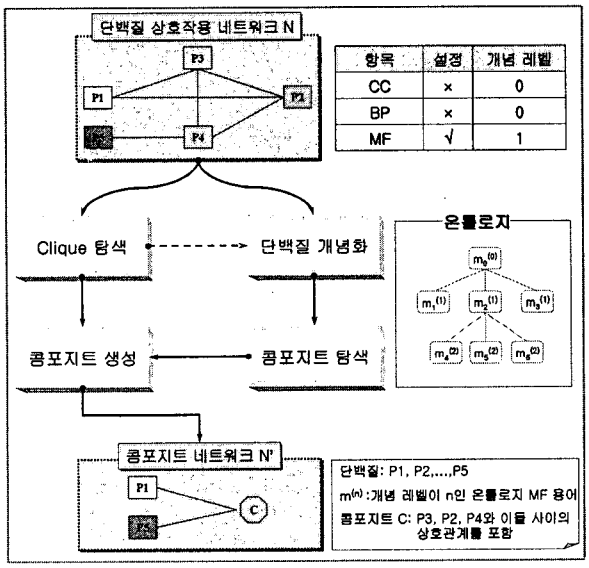
본 논문에서는 이 단점을 보완하기 위해 개념적으로 인접한 부분 네트워크를 탐색하고, 이들을 선택적으로 하나의 콤포지트로 추상화함으로써 복잡한 내트워크를 개념적으로 단순화할 수 있은 방법을 제안한다.

2. 네트워크 추상화

단백질 상호작용 네트워크에서 바이오 콤포지트를 개념 기반 방식으로 탐색하여 추상화하기 위해서는 네트워크에 포함된 단

백질들을 일정한 표준 통제 용어(Controlled Vocabulary)들로 개념화해야 한다. 또한, 이 통제 용어들 사이의 의미적 관계를 정의한 온톨로지가 요구된다. 특정 용어가 다른 용어들과 의미적인 관계를 가질 때, 이 용어는 온톨로지 내에서 하나의 개념을 형성하게 된다. 또한, 이 개념은 다른 개념들과의 의미적 상하위 관계(IS-A)에 의해 특정 개념 레벨을 가진다.

한편, 공개된 단백질들에 관한 많은 정보들을 포함하고 있는 데이터베이스 Swiss-Prot은 단백질 각각에 여러 특성들을 유전자 온톨로지(GO) 용어들로 명세하고 있다. 또한, GO는 통제 용어들을 3 가지 범주(BP: Biological Process, CC: Cellular Component, MF: Molecular Function)에 따라 분류하고, 이들 사이의 의미적 관계를 계층적으로 표현하고 있다. 따라서, 본 논문은 Swiss-Prot의 단백질 데이터베이스와 GO를 사용하였다.



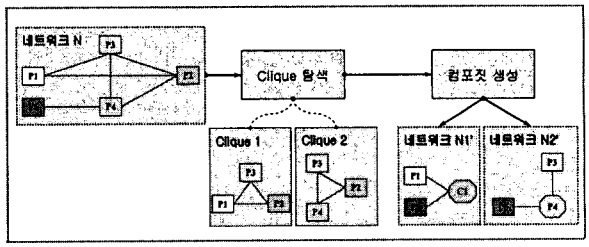
[그림 1] 네트워크 추상화 과정

[그림 1]은 콤포지트 탐색을 통해 네트워크를 추상화하는 과정을 나타내고 있다. 이때, 클릭에 의하거나 단백질의 개념을 통해 네트워크에서 의미적으로 연관성이 있는 여러 부분 네트워크를 탐색한다. 이 부분 네트워크들 중 의미가 있는 것들은 콤포지트로 추상화된다. 따라서, 하나의 콤포지트는 단백질들 사이의 관계로 구성된 하나의 부분 네트워크의 추상화된 단위로 간주될 수 있다. 여기서, 콤포지트 생성 모듈은 네트워크에서 탐색된 부분 네트워크를 콤포지트로 추상화하여 네트워크를 재구성한다.

네트워크에서 콤포지트들은 3 가지 방법에 의해 탐색된다. 즉, 부분 네트워크가 클릭 구조를 가지는 콤포지트, 하나의 개념을 공통으로 가지는 콤포지트 그리고 클릭 중에서 하나의 개념을 공통으로 가지는 콤포지트들이 탐색될 수 있다. 다음은 이 방법들을 예를 통해 설명한다.

3. 바이오 콤포지트 탐색

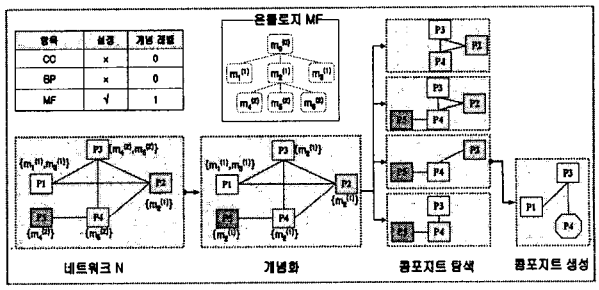
부분 네트워크를 구성하는 노드가 다른 모든 노드들과 관계를 가질 때, 이 부분 네트워크는 완전한 구조의 클릭을 형성하게 된다. 네트워크에서 이 클릭은 하나의 클러스터에 클릭 구조를 유지하는 인접한 노드를 계속 추가함으로써 탐색될 수 있다.



[그림 2] 클릭 콤포지트 탐색

[그림 2]에서 {P1}를 중심으로 클러스터 {P1,P2}, {P1,P3}를 생성할 수 있다. 또, {P1,P2}에서 {P1,P2,P3}와 {P1,P2,P4}를 생성할 수 있는데, {P1,P2,P4}은 클릭 구조를 이루지 않기 때문에 클러스터에서 제외시킨다. 이 과정을 각각의 노드에 대해 반복하면 3개 이상의 노드를 가지는 클릭을 탐색할 수 있다. 이들 중 의미있는 클릭을 콤포지트로 생성하여 네트워크를 재구성할 수 있다. 본 논문은 빠른 클릭 탐색을 위해 [2]을 이용하였다.

단백질 상호작용 네트워크에서 콤포지트로 표현될 수 있는 대표적인 것으로 단백질 콤플렉스가 있다. 콤플렉스에 포함된 단백질들은 서로 연합하여 특정한 생물학적 역할을 수행하기 때문에 단백질들 사이의 서로 밀접한 의미 관계를 가지고 있다. 따라서, 이런 콤플렉스를 탐색하기 위해서는 특정 개념을 중심으로 단백질들 사이에 상호 유사한 관계를 가지고 있는 부분 네트워크를 탐색할 수 있어야 한다.



[그림 3] 개념 콤포지트 탐색

[그림 3]은 온톨로지를 이용하여 단백질들 사이에 개념적으로 유사한 관계를 가지는 부분 네트워크를 N에서 탐색하는 과정을 설명하고 있다. 먼저, 사용자는 단백질을 개념화하기 위해 GO의 여러 범주를 설정할 수 있지만 예에서는 범주(MF)와 개념 레벨(1)을 설정하여 설명한다. 이 설정에 따라 N의 각각의 단백질(예,

P3)은 MF의 온톨로지 용어({$m4^{(2)}$,$m5^{(2)}$})로 대치된다. 다음으로, 개념화 모듈에서는 이 용어들을 개념 레벨의 용어 {$m2^{(1)}$}로 매핑한다.

개념 콤포지트는 네트워크에서 개념 클러스터에 의미적으로 밀접한 개념을 가지는 인접한 노드를 계속 추가함으로써 탐색될 수 있다. 즉, {P3}를 중심으로 P2와 P4이 P3의 개념 $m2^{(1)}$을 가지고 있기 때문에 클러스터 {P2,P3}, {P3,P4}를 생성할 수 있다. 또한, 같은 방법으로 {P2,P3}에서 {P2,P3,P4}으로, 다시 {P2,P3,P4,P5}로 클러스터를 확장할 수 있다. 이 과정을 반복하여 3개 이상의 단백질들을 가지는 부분 네트워크 4개를 탐색할 수 있다. 사용자는 이 부분 네트워크들 중 특정한 의미를 가지는 것만을 선택하여 콤포지트로 추상화할 수 있다.

이 개념 콤포지트 탐색 방법을 온톨로지의 CC 범주와 동시에 사용할 경우, extra cellular, membrane, inter cellular, nucleus 등으로 네트워크를 구분하여 추상화할 수 있다. 또한, 클릭을 먼저 탐색하고 이 클릭 중 개념적으로 밀접한 관계를 가지는 클릭만을 위와 같은 방법으로 탐색하여 네트워크를 추상화할 수 있다.

4. 구현

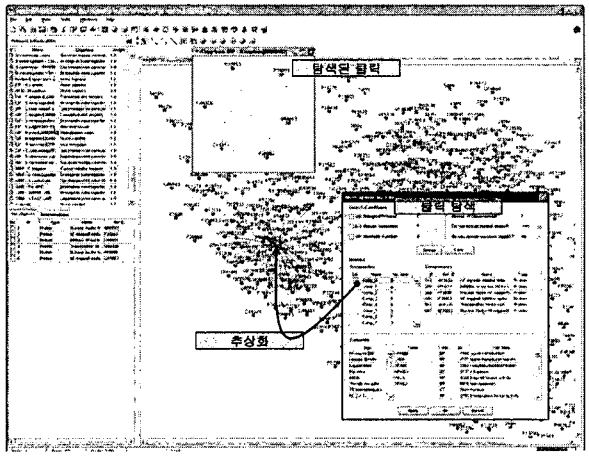
본 논문에서 제안하는 추상화 방법은 단백질 상호작용 네트워크 관리 시스템의 하나의 기능으로 구현되었다. 구현 환경은 JAVA 언어와 Oracle 데이터베이스를 이용하였다. 구현된 추상화 방법은 DIP에서 제공하는 2005년 1월 데이터를 이용하여 테스트되었다. 이 네트워크 데이터는 Homo Sapience에 대해 1,054개의 단백질과 1,344개의 상호작용 관계로 구성되어 있다.

[그림 4]는 클릭을 탐색하고 이중 하나를 추상화한 예를 보여주고 있다. 클릭 콤포지트는 371개가 탐색되었으며, 최대 클릭은 6개의 단백질로 구성되어 있다. 여기서, 온톨로지 범주를 BP로 하고 개념 레벨을 3으로 하였을 경우 총 16개의 개념 클릭들이 탐색된다. 이중 최대 클릭의 단백질 개수는 4개이다.

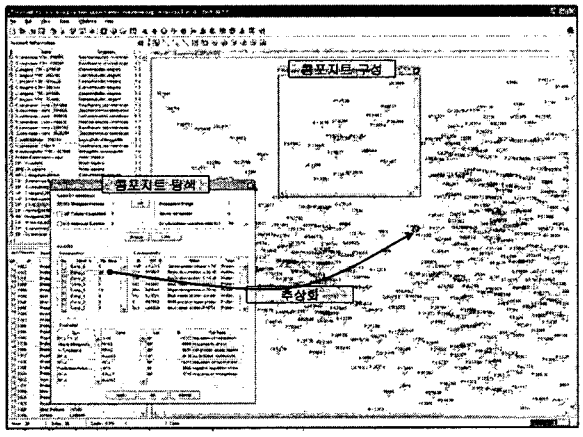
[그림 5]는 온톨로지 범주 BP와 개념 레벨 3을 통해 개념 콤포지트들을 탐색하고 이중 하나를 콤포지트로 추상화한 예이다. 총 73개의 개념 콤포지트가 탐색되었으며 최대 컴포지트는 20개의 단백질로 구성되어 있다. 이 단백질은 대부분 DNA 복제와 관련되어 있다. 위 방법들로 탐색된 콤포지트들은 서로 같은 단백질을 공유할 수 있으며, 하나의 콤포지트로 추상화된 노드는 다시 처음 상태로 확장될 수 있다. 또한, 콤포지트를 가지고 있는 네트워크에 대해 다시 추상화를 수행할 수도 있다.

5. 결론

본 논문은 단백질 상호작용 네트워크을 생물학적인 의미 연관성을 가지는 콤포지트를 탐색하여 추상화하는 방법을 설계하고 구현하였다. 이 추상화는 네트워크에 포함된 구조적으로 완전한 클릭, 개념적으로 인접한 부분 네트워크 그리고 개념적으로 인접한 클릭을 탐색하여 콤포지트로 추상화할 수 있다. 특히, 콤포지트 탐색은 방대한 네트워크에서 특정한 의미를 가지는 부분 네트워크를 탐색하고 분석하는데 매우 유용하게 이용될 수 있다.



[그림 4] 클릭 콤포지트 탐색 화면



[그림 5] 개념 콤포지트 탐색 화면

참고문헌

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

[2] I.M. Bomze, The Maximum Clique Problem. Handbook of Combinatorial Optimization, Vol. 4. 1999

[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.