단백질 상호작용 네트워크 항해를 위한 시스템

최재훈°, 박선희
한국전자통신연구원
{jhchoi°,shp}@etri.re.kr

A System Architecture for Navigating Protein Interaction Networks

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

요 약

본 논문에서는 생물체의 세포에 존재하는 방대한 단백질들 사이의 복잡한 관계들로 표현되는 상호작용 네트워크를 효율적으로 항해할 수 있는 시스템을 제안한다. 이 항해 시스템은 네트워크 검색 컴포넌트 그리고 상호작용 정보 시각화 컴포넌트로 구성된다. 네트워크 검색 컴포넌트는 사용자 질의를 통해 여러 네트워크들 중에서 사용자가 관심이 있는 네트워크들만을 개념기반으로 검색할 수 있도록 지원한다. 또한, 사용자는 시각화 컴포넌트를 통해 검색된 하나의 네트워크에 포함된 복잡한 노드들 사이의 관계 정보를 자동으로 시각화할 수 있다.

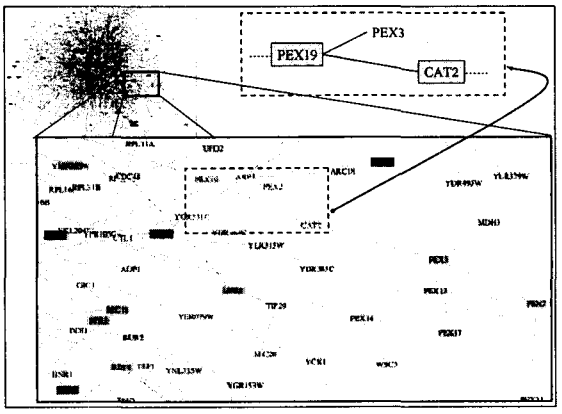
1. 서론

휴먼 지놈 프로젝트를 통해 인간의 염색체 24쌍에 대한 대략적인 염기 서열이 밝혀졌다. 염색체에는 후손에게 유전되는 유전자를 포함하고 있으며, 이 유전자는 일반적으로 세포에서 단백질로 번역된다. 단백질은 세포의 특정한 구성요소로서 몇 개의 생물학적 역할에 참여하며 고유의 분자 기능을 수행하게 된다.

일반적으로 하나의 단백질은 고유의 기능을 가지고 있지만, 생체 내에서 특정한 생물학적 역할을 하기 위해서는 여러 다른 단백질들과의 상호작용을 하게 된다. 따라서, 단백질들 사이의 생물학적인 상호작용들을 여러 관계로 표현할 수 있다. 이 단백질 상호작용 관계 정보는 기본적으로 이스트 투 하이브리드(yeast two hybrid) 라는 생물학적 실험을 통해 추출된다[1][6]. 현재, 이 실험을 통해 구축된 단백질 상호작용 관계 정보는 생물 종에 따라 분류하여 데이터베이스에 체계적으로 관리고 있으며, 그 일부는 공개되고 있다. 대표적인 데이터베이스로 YPD(Yeast Proteome Database), PIMdb(Drosophila Protein Interaction Map database), BIND(Biological Interaction Network Database), DIP(Database of Interacting Protein) 등이 있다. 하나의 세포 내에는 많은 단백질들 사이의 상호작용 관계들이 존재하며, 이들은 그래프 형태의 관계 네트워크로 명시될 수 있다. 즉, 단백질은 노드, 단백질 사이의 관계는 링크로 표현될 수 있다[3][4].

예를 들어, [그림 1]은 [5]에서 제시한 2358개의 효모(yeast) 단백질들에 대한 상호작용 관계 네트워크를 나타내고 있다. 여기서, 효모가 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.

본 논문에서는 생물체의 세포에 존재하는 방대한 단백질들 사이의 복잡한 관계들로 표현되는 상호작용 네트워크를 효율적으로 항해할 수 있는 시스템을 제안한다. 이 항해 시스템은 네트워크 검색 컴포넌트와 상호작용 정보 시각화 컴포넌트로 구성된다. 네트워크 검색 컴포넌트는 사용자 질의를 통해 여러 네트워크들 중에서 사용자가 관심이 있는 네트워크들만을 개념기반으로 검색할 수 있도록 지원한다. 또한, 사용자는 시각화 컴포넌트를 통해 검색된 네트워크에 포함된 복잡한 노드들 사이의 관계 정보를 효율적으로 시각화할 수 있다.

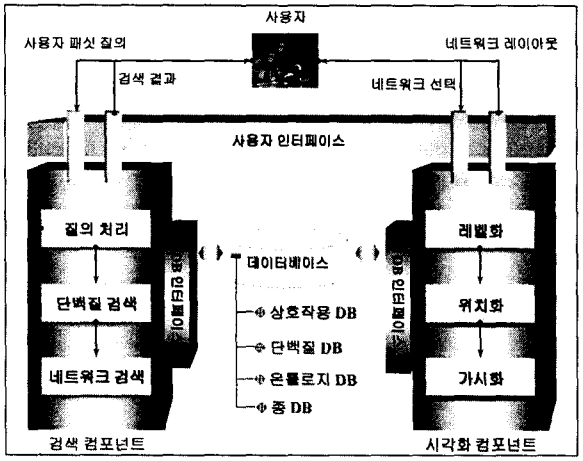


[그림 1] 단백질 상호작용 관계 네트워크의 예

2. 단백질 상호작용 네트워크 항해 시스템

일반적으로 단백질 상호작용 관계 네트워크는 매우 방대한 단백질 노드들 사이의 대한 복잡한 관계들로 구성되어 있다. 또한, 하나의 기능을 표현하는 네트워크가 생물체의 종에 따라 참여하는 단백질뿐만 아니라 이들 사이의 관계가 다를 수 있다. 또한, 같은 종에 대해 여러 종류의 네트워크가 존재할 수 있다. 따라서, 단백질 상호작용 네트워크를 위한 시스템은 데이터

베이스로부터 자신이 원하는 종의 특정 네트워크를 선별적으로 검색할 수 있어야 한다. 또한, 복잡한 관계들로 표현된 네트워크를 최적화된 형태로 시각화할 수 있어야 한다. 이를 위해 본 논문에서는 [그림 2]와 같은 단백질 상호작용 네트워크 항해 시스템을 제안한다.



[그림 2] 네트워크 항해 시스템 구성도

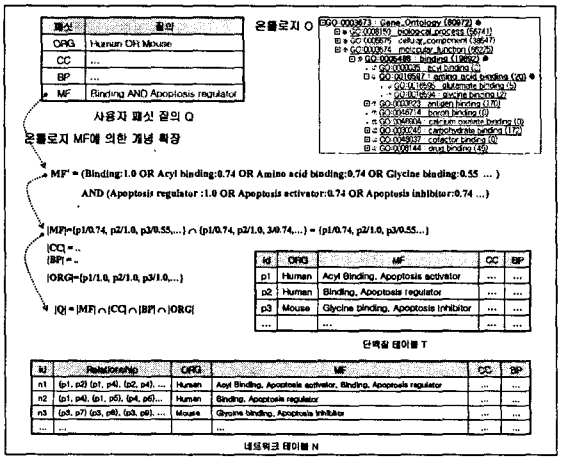
이 시스템은 크게 네트워크를 검색하고 시각화할 수 있는 2개의 컴포넌트로 구성되어 있다. 또한, 상호작용 관계, 단백질, 온톨로지 그리고 종에 대한 데이터베이스를 가지고 있다. 검색 컴포넌트는 사용자 질의를 만족하는 단백질을 검색한 다음, 이 단백질들을 포함하고 있는 네트워크를 다시 검색한다. 또한, 사용자는 선택적으로 사용자 질의로부터 직접 네트워크를 검색할 수 있다. 이때, 사용자 질의는 "종", "세포 구성 요소", "생물학적 역할" 그리고 "분자 기능"에 대한 각각의 패싯으로 표현된다.

시각화 컴포넌트는 상호작용 관계 네트워크를 여러 단계로 레벨화한 다음, 각각의 레벨에 포함된 노드들에 대한 위치 좌표를 구한다. 이때, 특정 레벨의 위치 데이터는 이전 단계의 위치 데이터를 기반으로 구하게 된다. 이렇게 구해진 단백질 상호작용 네트워크의 노드 좌표를 사용자 인터페이스에 적절히 위치시키고 이들에 대한 링크들과 함께 네트워크를 가시화한다.

현재, Biolayout, PIMRider, PIVOT(Protein Interactions VisualizatiOn Tool), InterView 등과 같이 네트워크를 자동으로 시각화해 주는 많은 도구들이 개발되어 있다. 그러나, 이들은 FDP(Force-Directed Placement)[3]을 이용하고 있으나, 본 시각화 컴포넌트는 보다 빠른 가시화를 위해 네트워크를 여러 단계로 레벨화는 MFDP(Multilevel algorithm for Foce-Directed Placement)[7]를 이용하였다.

3. 네트워크 검색

본 논문은 다음 2 가지 방법을 통해 단백질 상호작용 네트워크 검색한다. 첫째 방법은 먼저 패싯 질의를 통해 단백질을 검색한 다음, 사용자가 이들 중 적당한 단백질을 선택하면 이 선택된 단백질들을 포함하고 있는 네트워크를 검색한다. 둘째 방법은 중간의 단백질을 검색하는 단계 없이 바로 패싯 질의를 통해 네트워크를 검색하는 방법이다. 이때, 유전자 온톨로지는 단백질의 3개 패싯에 대한 값들을 명세하기 위해 사용된다. 또한, 네트워크의 3 패싯은 이 네트워크가 포함하고 있는 단백질들의 패싯 값들을 조합함으로써 자동으로 명세된다.



[그림 3] 네트워크 검색 과정

[그림 3]은 네트워크 검색 과정을 예를 통해 설명하고 있다. 온톨로지(O)는 BP(Biological Process), CC(Cellular Component) 그리고 MF(Molecular Function) 3가지 관점으로 구성되어 있다. 사용자 패싯 질의(Q)는 ORG, BP, CC 그리고 MF 4개의 패싯으로 기술된다. 여기서, ORG(Organism)는 생물학적 종을 나타낸다. 각각의 패싯에 대한 질의는 불리언 연산자의 조합으로 표현된다.

예를 들어, Q.ORG="Mouse OR Human" 그리고 Q.MF="Binding AND Apoptosis regulator"이고 다른 패싯은 무조건 항 패싯이라고 가정하자. 이 질의 Q.MF에 대해서는 온톨로지에 의한 개념확장이 수행된다. 즉, Q.MF에 포함된 하나의 개념 'Binding'에 대해 온톨로지의 하위 개념들에 의해 "Binding:1.0 OR Acyl binding:0.74 OR Amino acid binding:0.74 OR Glycine binding:0.55"으로 확장된다. 이때, o1= 'Binding'과 확장된 개념 o2 사이의 가중치는 아래 식에 의해 계산된다.

$$Ontology(o_1, o_2) = e^{-0.3 \times DIST(o_1, o_2)}$$

이때, 가중치는 0과 1사이의 퍼지 값으로 간주하며, ∩과 ∪은 퍼지 교집합과 합집합을 나타낸다. 또한, 두 퍼지 값에 대한 ∩과 ∪의 평가는 가장 기본적인 퍼지 함수 min과 max를 이용한다.

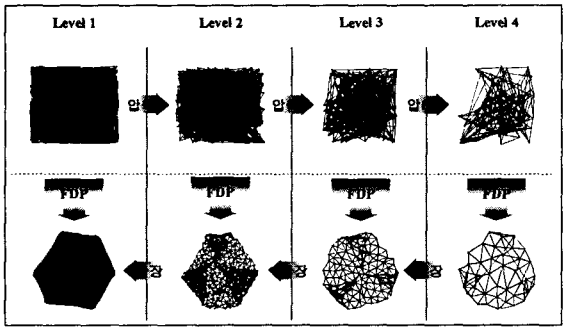
단백질 테이블 P에서 하나의 Mouse 단백질 p3는 Q.MF의 'Glycine binding:0.55'과 'Apoptosis inhibitor:0.74'에 의해 두 가중치의 최소값인 0.55으로 검색된다. 검색된 단백질 중 사용자가 p1과 p3를 선택하여 이 두 단백질을 동시에 포함하고 있는 네트워크 n1를 검색할 수 있다.

네트워크 테이블 N에 대해서도 위의 패싯 질의를 통해 같은 방법으로 검색

할 수 있다. 이때, 각 네트워크에 부여된 MF, CC 그리고 BP에 대한 패싯 값들은 그 네트워크를 구성하는 단백질들이 가지는 패싯 값들의 조합으로 자동 구축될 수 있다. 예를 들어, 네트워크 n1은 단백질 p1과 p2를 포함하고 있기 때문에 n1.MF=p1.MF ∪ p2.MF .....으로 n1의 MP 패싯 값들을 자동 구축할 수 있다.

4. 네트워크 시각화

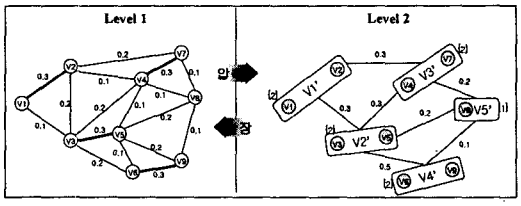
단백질 상호작용 관계 네트워크는 매우 많은 노드와 링크로 구성되어 있어 시각화하기 매우 어렵다. 일반적으로 노드들의 위치를 계산하기 위해 FDP 알고리즘 [3]을 많이 사용하고 있다. FDP는 각 노드들에 대해 전역 계산(global force calculation), 지역 계산(local force calculation) 그리고 재배치 연산(reposition)을 반복 수행한다. 하나의 노드에 대해 전역 계산은 다른 모든 노드들과의 상관 관계를 고려하며, 지역 계산은 이 노드와 인접한 노드들과의 상관관계만을 고려한다. 이때, 두 노드 사이에 자연 거리 k를 상수 값으로 처리한다. 이 방법은 매우 단순하며 일반적인 장점을 가지는 반면, 매우 많은 노드와 링크들을 가지는 단백질 상호작용 관계 네트워크에 대한 위치화에는 많은 시간이 요구되는 단점을 가지고 있다. 이 단점을 해결하기 위해 본 논문에서는 MFDP [7]를 이용하였다.



[그림4] 다중 레벨 네트워크 시각화 과정

[그림 4]는 MFDP에 의한 시각화 과정을 설명하고 있다. 첫째, 각 레벨의 네트워크의 노드 수가 임계값 이하가 될 때까지 압축을 한다. 둘째, 마지막 레벨에 대한 FDP 알고리즘을 수행하여 이 레벨의 노드들에 대한 위치 좌표들을 계산한다. 셋째, 이전 레벨로 네트워크를 확장한다. 이때, 이전 레벨에서 계산된 위치 정보는 확장된 네트워크의 노드에 그대로 반영한다. 이 위치 정보를 바탕으로 다시 FDP를 수행하여 현재 레벨의 노드 위치를 계산한다. 이 과정을 반복하여 초기 네트워크까지 확장함으로써 각 노드에 대한 위치를 계산하게 된다. 즉, 이 방법은 이전 레벨의 위치 정보를 이용하여 현재 레벨의 위치정보를 계산함으로 복잡한 네트워크의 노드들에 대한 위치 정보를 빠르게 계산할 수 있다는 장점을 가지고 있다.

[그림 5]는 네트워크 레벨화 과정을 예를 통해 설명하고 있다. 즉, 하나의 노드 v1과 관련 정도가 높은 링크에 있는 노드 v2가 다음 레벨의 노드 v1'가 된다. 마찬가지로 v3과 v5 역시 하나의 노드 v2'로 표현된다. 또한, (v1.v3,0.1)과 (v2,v3,0.2)가 존재하기 때문에 (v1', v2', w=0.1+0.2)가 존재하게 된다. 또한, v1'에 대한 압축 정도는 2가 된다. 다음 압축과정에서 이 값이 높을수록 낮은 우선 순위를 갖게 된다. 이 과정을 네트워크 노드 수가 임계값 이하일 때까지 수행한다. 이때, Level 1에 대해 $k_1=0.75*k_2$이고 마지막 레벨 L에 대해 k_L=(링크 차수 합)/(링크 총 개수)로 계산된다.



[그림 5] 네트워크 레벨화 과정

5. 결론 및 향후 연구

생물학적 관점에서 매우 중요한 정보를 포함하고 있는 단백질 상호작용 관계들은 복잡한 네트워크로 표현될 수 있다. 본 논문에서는 이 상호작용 네트워크를 효율적으로 항해할 수 있는 시스템을 제안하였다. 특히, 네트워크 검색 컴포넌트는 패싯 질의를 통해 여러 네트워크들 중에서 사용자가 관심이 있는 네트워크들만을 개념기반으로 검색할 수 있다. 이때, 개념 기반 질의 처리를 위해 유전자 온톨로지를 이용하였다. 이 검색된 네트워크는 시각화 컴포넌트를 통해 효율적으로 시각화할 수 있도록 설계하였다. 이 시각화를 위해 MFDP 알고리즘을 이용하였다. 또한, 이 시스템은 시각화된 네트워크의 각각의 노드 및 관계에 대한 자세한 생물학적 정보를 인터넷을 통해 참조할 수 있도록 지원한다.

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] T. M. J. Fruchterman and E. M. Reingold, "Graph Drawing by Force-Directed Placement," Software: Practice and Experience, Vol. 21, No. 11, 1991.

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

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

[5] C. Walshaw, "A Multilevel Algorithm for Force-Directed Graph Drawing," Graph Drawing 8th Intl. Symp, Berlin, 2001. [2] S. Field, and O. Song, "A Novel

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