단백질 상호작용 네트워크의 개념 분류 레이아웃

방선이° 최재훈 박종민 박수준
한국전자통신연구원
{ slbang°, jhchoi, jmpark93, psj}@etri.re.kr

Conceptual Classification Layout of Protein-Protein Interaction Networks

SunLee Bang°, JaeHun Choi, JongMin Park, SooJun Park
Electronics & Telecommunications Research Institute(ETRI)

요 약

본 논문은 온톨로지를 이용하여 단백질 상호작용 네트워크를 개념적으로 분류하여 레이아웃하는 방법을 제안한다. 상호작용 네트워크를 이루는 단백질은 온톨로지의 표준 통제 용어에 대한 주석 정보를 가지고 있으므로 동일 분류에 해당하는 통제 용어를 가지고 있는 단백질들은 근접한 곳에 위치하도록 레이아웃한다. 이는 기존 물리적 레이아웃에 기능별 그룹화를 해줌으로써 복잡한 네트워크를 개념적으로 분석할 수 있도록 한다. 또한, 동일 분류에 속하는 단백질들을 한 노드로 대응하여 레이아웃 알고리즘을 수행함으로써 기존의 그래프 표현 알고리즘 보다 빠르게 시각화할 수 있다.

1. 서론

일반적으로 하나의 단백질은 고유한 기능을 가지고 있지만, 생체 내에서 특정한 생물학적 역할을 하기 위해서 여러 다른 단백질들과 다양한 상호작용을 한다. 따라서, 하나의 세포 내에는 많은 단백질들 사이에 복잡한 상호작용 관계들이 존재한다. 현재, 대부분의 단백질간의 상호작용 데이터는 'Yeast Two-Hybrid' 와 'co-AP/MS'라는 생물학적 실험을 통해 빠르게 추출 되고 있으며, 추출된 데이터는 BIND(Biomolecular Interaction Network Database), DIP(Database of Interacting Protein), IntAct 등과 같이 데이터베이스에 체계적으로 관리되고 있다[1].

이러한 단백질들간의 상호작용 데이터를 단백질은 노드로 이들 사이의 상호작용을 에지로 표현하면 네트워크로 나타낼 수 있다[2]. 방대한 단백질들 사이의 복잡한 관계들로부터 특정 단백질이 아닌 전체적인 생체 메커니즘을 이해하기 위해 네트워크를 분석하기 위한 응용 시스템의 연구가 요구되고 있다. 네트워크 분석은 네트워크를 표현하는 레이아웃에 정보들이 얼마나 효율적으로 표현되었느냐에 따라 네트워크를 개념적으로 파악할 수 있도록 표현을 해주어야 한다. 서로 관계를 가지고 있는 대용량의 데이터를 쉽게 이해하기 위해 그래프로 표현하는 방법이 많이 사용되고 있다. 그러나 대부분의 그래프 표현방식은 데이터들간의 관계에만 의존한 물리적인 표현만을 하고 있다. 네트워크 크기가 방대하므로 효과적인 분석을 위해 단백질 간의 물리적 관계뿐만 아니라 단백질의 정보를 고려한 시각화 방법이 요구된다.

2. 관련연구

단백질 상호작용 네트워크를 시각화하기 위해 'Force-Directed Placement'(FDP) 알고리즘이 많이 사용되고 있다[2]. 이 알고리즘은 노드와 에지의 집합에 대해 force를 지정하여 균형 상태를 이루도록 레이아웃한다. 에지들이 겹쳐져 레이아웃되는 것을 방지하기 위해, 연결된 노드들간의 에지는 서로 당기는 local force로 보며, 연결되지 않은 노드들은 서로 밀어내는 global force로 본다. 이 알고리즘은 융통성 있고 구현하기 쉬우며 드로잉 결과도 양호하기 때문에 많이 사용되지만, 대용량의 데이터에 대해서는 느리다는 단점이 있다. 속도향상을 위해 노드집합에 대해 다단계로 클러스터를 이룬 후, 이를 확장하는 과정에서 FDP를 적용하는 'Multilevel for Force-Directed Placement' (MFDP) 알고리즘이 있다[3]. 그러나, 단계별로 시작노드를 랜덤하게 설정하며 각 반복 단계에서 모든 쌍의 노드들 간의 force를 계산해야 하기 때문에 허브노드와 같이 한 노드에 이웃노드들을 많이 가지고 있는 경우는 이를 처리하는

데 시간이 많이 소요된다.

기존의 단백질 상호작용 네트워크를 지원하는 시스템들의 시각화 방법은 대부분 물리적인 관계만을 기반으로 표현하고 있다. 따라서, 대용량의 단백질에 대해 다수의 관계를 포함하고 있는 네트워크를 파악하기 위해서는 단백질의 정보를 함께 표현하는 시각화가 연구되고 있다. 실제 단백질 상호작용 네트워크를 FDP 관련 알고리즘을 기반으로 시각화하며 단백질의 정보를 함께 적용하여 시각화해 주는 시스템에는 Osprey 등이 있다. Osprey[4]는 단백질 노드에 주석되어있는 온톨로지 용어에 대해 색상을 달리하여 표현하고 있으며 기능 클러스터링을 적용하여 같은 온톨로지 용어를 가지는 노드들을 클러스터링 하여 보여준다. 이는 한꺼번에 노드의 정보와 관계를 파악할 수 있다는 장점이 있지만, 기존 물리적인 관계를 기반으로 하지 않고 블록별로 클러스터링 되어 표현되고 있어, 많은 수의 단백질을 기반으로 하여 네트워크가 구성될 때는 노드집합 전체에 대해 관계를 한눈에 파악하기가 어렵다.

따라서, 본 논문에서는 단백질에 주석되어있는 온톨로지의 용어 정보를 이용하여 동일 기능을 수행하는 단백질들을 서로 가까운 곳에 배치하도록 시각화 알고리즘을 재구성하고자 한다. 이는 MFDP 알고리즘을 통해 단백질간의 관계를 물리적으로 보여주고 있으나 같은 분류의 단백질들을 클러스터링하여 보여줌으로써 물리적인 시각화에서 개념적 시각화로의 전환으로 보다 빠르게 네트워크를 파악할 수 있다.

3. 개선된 MFDP 알고리즘

MFDP 알고리즘은 랜덤하게 한 노드를 선정한 다음, 연결된 다른 한 노드와의 합병 여부를 계산한 후, FDP 알고리즘으로 확장하여 시각화한다. 분류 알고리즘을 기반으로 같은 분류에 속하는 노드들을 한꺼번에 합병하기 위해 기존 MFDP알고리즘에 대해 시작노드 선정단계, 합병단계, 확장단계를 다음과 같이 개선한다.

합병과정을 수행하기 위해 기준이 되는 시작노드 선정은 다음과 같이 순위화된다. 네트워크상에서 다른 단백질들과 많은 관계를 가지고 있는 단백질일수록 중요성을 보이므로 연결된 그래프에서 이웃노드가 가장 많은 노드를 시작노드로 선정한다. 이웃노드의 수가 같다면, 한 노드를 기준으로 클러스터링 되는 것을 방지하기 위해, 이웃하는 노드의 차수와의 합이 가장 작은 노드를 선정한다. 각 단계별로 연결된 그래프의 시작노드를 중심으로 연결된 모든 노드들을 리스트에 저장하여 합병된 새로운 노드로 대치하여 최종 그래프의 노드가 1개 혹은 3개를 이룰 때까지 합병한다. 그림 1의 그래프는 이웃노드의 수가 가장 많은 1번 노드를 시작으로 이웃하는 2,3,4,5번 노드를 합병한다. 다음단계에선 합병된 노드의 이웃노드 개수가 가장 많으므로 6,7,8번 노드를 합병한다. 여러 노드를 한꺼번에 합병함으로써 기존 MFDP 알고리즘인 경우 총 3~5단계가 나오는 과정을 총 2단계를 거쳐 마칠 수 있다.

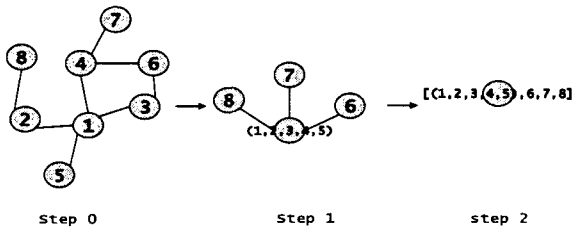

그림 1 합병 과정

합병노드들을 확장하기 위해 기존 MFDP 알고리즘의 초기 위치를 잡는 방법을 다음과 같이 개선한다. 최종합병단계에서 기준이 되는 노드를 중심으로 합병된 노드들을 자연거리상수(natural spring force)를 지름으로 하는 원주 위에 균일하게 위치하도록 한다. 합병된 노드들은 기준노드의 y좌표와 동일한 위치를 시작으로 합병된 노드수로 분할하여 위치시킨다. 이미 전 단계에서 확장된 노드와 관계를 가지는 노드는 관계가 있는 노드와 가까운 분할 포인트에 위치시키고, 관계가 없는 노드는 빈 분할 포인트에 위치시키도록 한다. 그림 1의 최종합병그래프를 확장하면 그림 2와 같다.

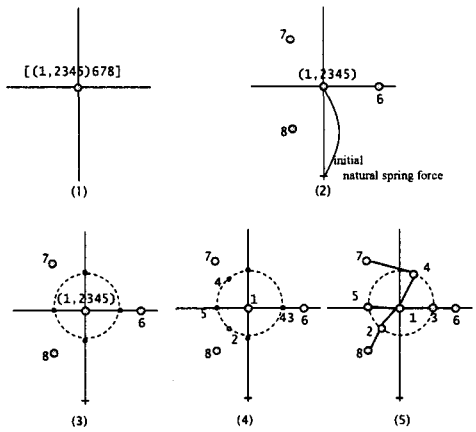

그림 2 확장 단계에서 노드의 초기위치 설정

4. 온톨로지를 이용한 개념 분류

단백질 상호작용 네트워크를 개념적으로 파악할 수 있도록 기존의 물리적인 표현 구조는 유지하면서 같은 기능을 하는 단백질

들은 서로 가까운 곳에 위치시킨다. 네트워크의 단백질 노드는 GO[5]의 세가지 계층구조(BP: Biological Process, CC: Cellular Component, MF: Molecular Function)를 이루는 용어로 주석되어있다. 기능별 분류를 위해 BP와 MF범주를 기반으로 단백질 노드의 분류를 수행한다. 개념적으로 분류하여 관련 단백질들을 클러스터링하여 시각화하는 방법은 기존의 물리적 구조의 표현 방법에 비해 네트워크를 보다 개념적으로 파악할 수 있다.

MFDP 알고리즘에서 먼저 합병된 노드들이 근처에 위치하게 되므로 GO의 각 범주의 동일 계층구조 상에 있는 단백질 노드들을 합병한다. GO의 특정 범주에 대해 개념화를 하기 위해 레벨을 준 후, 각 노드들을 해당 레벨의 GO용어로 변환한다. 합병과정을 수행하기 위해 시작노드는 다음과 같은 조건으로 확장하여 순위화한다. 물리적인 구조를 먼저 고려하여 이웃노드의 수가 가장 많은 노드, 그 노드가 GO용어로 주석이 되어있지 않다면, 다음으로 이웃노드의 수가 많은 노드를 설정한다. 이웃노드의 수가 동일한 노드가 두 개 이상 존재할 시는 GO용어를 보다 많이 가지고 있는 노드를 우선시 한다. GO 용어의 수가 동일한 경우는 노드들이 한쪽으로만 치우쳐 합병되는 것을 막기 위해 이웃노드와의 차수의 합이 작은 노드를 시작노드로 선정한다.

단백질 노드는 한 범주 내 여러 GO용어로 주석되어 있으므로 시작노드의 GO용어와 이웃노드들 중 GO용어의 교집합이 존재하는 노드들에 대해 합병을 수행한다.

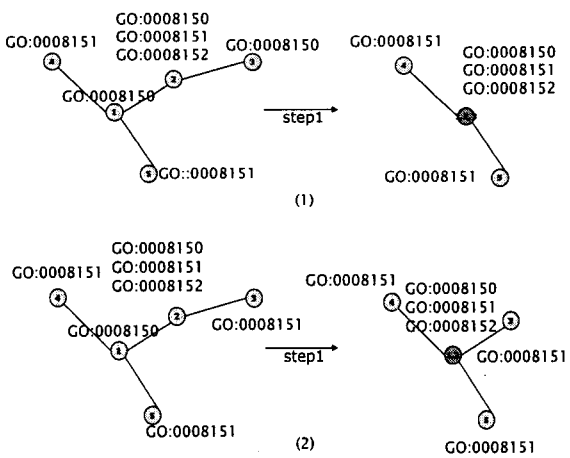

그림 3 GO용어에 대한 단백질 노드 분류

단계별로 합병되어 새로 생성된 노드의 GO용어는 기존 정보를 유지하기 위해 합집합을 이루는 GO용어로 주석한다. 그림 3의 (1)의 경우, 1번 노드를 시작으로 2번 노드의 GO:0008150에 대해 합병되고, 2번 노드와 이웃하는 3번 노드 역시 GO:0008150에 대해 합병된다. 합병된 노드에 대해선 1,2,3번 노드의 GO용어 합집합으로 주석한다. (2)의 경우, 1,2번 노드가 GO:0008150를 가지고 합병이 되지만, 3번 노드와는 교집합을 이루지 않으므로 1,2번 노드로 합병된 노드는 이들의 GO용어의 합집합으로 주석한다.

BP, MF 두 범주에 대해 함께 분류할 경우는 두 범주 모두에 대해 동일 용어를 가진 노드들은 1의 가중치를 한 범주에 대해 동일 용어를 가지는 경우는 교집합을 이루는 GO용어의 수에 대해 1보다 작은 가중치를 주며, 가중치는 FDP알고리즘 수행 시 force계산에 적용한다. 위의 과정을 통해 합병을 수행한 후, 모든 노드에 대해 합병할 노드가 없다면 레벨1에 이를 때까지 레벨을 낮추어 위의 과정을 반복 수행한다.

5. 결론 및 향후 연구 과제

본 논문에서 제시한 단백질 상호작용 네트워크의 개념 분류 레이아웃 방법은 기존 물리적 관계의 표현 위에 단백질노드에 주석되어 있는 GO 용어를 기반으로 같은 기능을 하는 단백질노드들을 분류하여 레이아웃함으로써 네트워크를 개념적으로 분석할 수 있으며, 기존 그래프 레이아웃 알고리즘에 비해 시각화 속도를 향상시켰다. 향후 과제로는 GO의 CC범주 계층구조를 반영하여 신호전달 시각화로 확장하고자 한다.

6. 참고 문헌

[1] P. Uetz and R. L. Jr. Finley, "From protein networks to biological systems," FEBS Letters, Vol. 579 No. 8, pp. 1821-1827, 2005.

[2] P. Uetz, T. Ideker and B. Schwikowski, "Visualization and Integration of Protein-Protein Interactions," E. Golemis, (ed.) Protein-Protein Interactions - A Molecular Cloning Manual. Cold Spring Harbor Laboratory Press, pp. 623-646.

[3] C. Walshaw, "A Multilevel Algorithm for Force-Directed Graph-Drawing," Journal of Graph Algorithms and Applications, Vol. 7, No. 3 pp. 253-285. 2003.

[4] B. J. Breitkreutz, C. Stark and M. Tyers, " Osprey: a network visualization system," Genome Biology, Vol. 4, Issue 3, Article R22, 2003.

[5] Gene Ontology, http://www.geneontology.org