대용량 단백질 상호관계의 시각화를 위한 기능기반 추상화 방법

김대희° , 최재훈 , 정재영 , 박선희
한국전자통신 연구원 바이오 정보팀
{dhkim98°, jhchoi, jjy72, sph}@etri.re.kr

A function-based abstraction method for visualizing the large scale of protein-protein interaction relationships

Dae-Hee Kim° , Jae Hun Choi , Jae Young Jung , Seon Hee Park
Bioinformatics Research Team, ETRI

요 약

이 논문은 대용량 단백질 상호작용의 관계를 효과적으로 시각화하기 위해 단백질이 가지고 있는 기능에 기반한 추상화 방법을 제안한다. 제안하는 방법은 FDP(force-directed placement) 알고리즘에 기반을 두고 있지만 다중 레벨 처리를 위해 기능에 기반한 추상화 방법과 확장을 사용한다는 점에서 차이점을 나타낸다. 제안하는 그래프 레이아웃 방법은 추상화, 위치화, 확장의 3부분으로 구성되어 있으며 특히 추상화 부분은 다중 레벨 처리를 포함한다.

1. 서 론

인간 유전체 초안이 발표되고 과학자들의 주요 연구는 유전자 한개나, 생화학적 경로에 관련되어 있는 유전자들을 분석하는 일이었다. 유전자가 만들어 내는 단백질의 기능 파악은 이 분야의 핵심적인 부분이다. 특히 단백질의 3차원 구조[1]를 알면 단백질의 기능을 억제하는 역할을 하는 신약개발에 몰두하고 있는 의료 화학자들의 연구에 큰 도움이 된다. 하지만 3차원 구조가 그 단백질의 기능에 관해서 별다른 정보를 주지 않을 때도 많다. 반면에 알려져 있는 다른 단백질과의 상호작용을 연구함으로써 중요한 단서를 얻을 수도 있다[2]. 단백질의 상호작용을 연구하긴 위한 기본 방법으로는 시각화를 들 수 있다.

그래프 레이아웃은 수학이나 전산과학 분야에서 광범위하게 사용되고 있다. 특히 생물학 분야에서는 단백질의 상호 작용과 관련하여 이를 시각화하기 위해 그래프 레이아웃이 사용된다. 생물학 분야에서 사용되는 단백질 상호작용 레이아웃은 두 가지 방법이 이용되고 있는데 첫째는 단백질을 그래프의 노드로 나타내고 상호작용을 간선으로 표현하는 방법이다[3][4]. 다른 접근 방법은 단백질의 기능을 정점으로 표현하고 그것들의 시맨틱 관계를 간선으로 표현하는 방법이다[5]. 일반적으로 전자가 이 분야에서 주로 연구되고 있다. 그러나 단백질 상호작용 데이터는 그 수의 방대함으로 인해서 설령 그래프로 시각화 하더라도 제대로 화면상에 표현하지 못하는 단점이 있기 때문에 그래프를 간결하게 보여주기 위한 방법이 필요하다. 게다가 단백질 상호작용의 추상화된 그래프는 생물학자들이 단백질 상호관계의 특별한 규칙을 찾아내는데 도움을 줄 수 있다.

일반적으로 단백질 상호작용 네트워크는 단백질과 그것들 사이의 관계로 표현되는 그래프로 정의되고 다음과 같이 표현할 수 있다. N=<P, R>. 여기서 P는 단백질들의 집합이고, R은 그들의 관계를 나타낸 집합이다. 앞서 설명한대로 이것은 그래프의 표현식인 G=<V, E> 의 형태와 유사한 것을 알 수 있다. 또한 단백질은 이름과 기능으로 표현 되어질 수 있기 때문에 다음과 같이 나타낼 수 있다. P=<n, F>. 여기서 n은 단백질의 이름이고, F는 단백질이 가지고 있는 기능들의 집합이다.

2. 추상화를 통한 그래프 시각화

이 장에서는 단백질 상호작용 관계를 효과적으로 시각화하기 위한 방법으로 기능기반 추상화 방법에 대해서 소개한다. 제안하는 방법은 3부분으로 구성되어 있고 각각의 부분은 다음과 같다.

각 부분을 설명하기에 앞서 다음의 표현식을 사용한다. 일반 적인 그래프를 나타내는 G=<V, E>로 나타낸다. 이것은 정점들의 집합 V, 간선들의 집합 E를 갖는 무향 그래프이다. 또한 G는 연결되어 있다고 가정한다. 임의의 정점 v에 대해서 Γ(v)는 정점 v에 인접한 정점들의 집합을 나타낸다. 즉 Γ(v)={ u∈V : (u, v) ∈ E}이다. 또한 |Γ(v)|는 정점 v 에 인접한 정점들의 수로 표현한다. 단백질 네트워크를 그래프의 형태에 맞게 표현하기 위해서 네트워크 N을 그래프인 G로, 단백질 집합인 P를 V로, 상호관계의 집합인 R을 E로 대응시키면 N=< P, F> 의 관계를 G=<V, E>로 나타낼 수 있다. 이하의 설명에서는 단백질을 정점, 그 상호관계를 간선으로 서술하며 생성

되는 중간 그래프 $G_i=<V_i, E_i>$로 나타낸다. 그러므로 $\Gamma(v, F)$를 다음과 같이 정의할 수 있다. 이것은 정점 v에 대해서 (여기서는 단백질이다) 기능 F를 가지고 있는 정점 v 에 인접한 정점들의 집합이고 $|\Gamma(v, F)|$는 그 수를 나타낸다. 여기서 |V|는 그래프가 가지고 있는 정점들의 수를 나타낸다.

2.1 기능 기반 추상화 방법

추상화 부분에서는 단백질이 가지고 있는 기능을 이용하여 상호작용 네트워크 그래프를 추상화 한다. 이 작업을 위해 단백질 기능에 관한 DB인 온톨로지를 이용한다. 이 데이터는 다음의 웹 사이트에서 다운로드 받을 수 있다. (www.geneontology.org)

그림1은 제안하는 기능 기반 그래프 시각화의 다이어그램 이다. 앞서 설명한대로 3부분으로 구성되어 있으며 추상화 이후에 정의된 레벨까지 반복적으로 위치화와 확장을 한다.

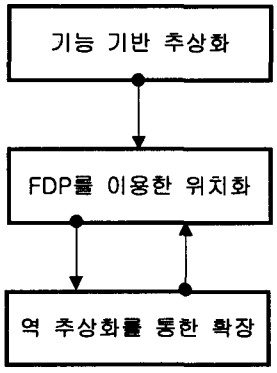

그림1. 기능 기반 그래프 시각화 다이어그램

다음의 표는 추상화 방법에 대한 순서를 설명하고 있다. 먼저 단계 1에서는 각각의 단백질에 대해서 단백질을 정점으로 상호관계를 간선으로 할당하고 단계 2에서는 하나의 정점에 대해서 주변의 정점들을 하나의 노드로 추상화(또는 그룹핑)하기 위해 기본적으로 매칭 방법을 사용한다. 각 정점들은 그 정점이 가지고 있는 기능과 같은 기능을 가지고 있고 상호관계가 있는 주위의 정점들과 매칭이 되어 하나의 클러스터를 형성하며 이 클러스터는 단계 3에 의해서 새로운 정점으로 할당된다. 단계 4에서는 생성된 새로운 정점과 간선으로 그래프가 정의되고, 사용자에 의해 조절된 최종 추상화의 크기를 결정하는 계수값과 그래프의 정점수 |V| 를 비교하여 일정 조건에 부합할 때까지 계속 반복한다(단계5). 그래프 매칭을 하는 구체적인 방법은 다음과 같다. 먼저 각 정점들에 대해서 $|\Gamma(v, F)|$의 차례 목록을 만들고 각각의 정점들을 방문하면서 하나의 정점에 대해서 $|\Gamma(v, F)|$의 값이 큰 것부터 단백질 기능을 참조하여 추상화를 한다. 이후 각각의 매칭된 정점들은 목록으로부터 사라진다. 즉 클러스터들은 상호관계가 있고 동시에 같은 기능을 갖는 단백질들을 추상화하게 된다.

추상화 순서
단계 1. 각각의 단백질에 대해서 단백질을 정점으로 상호관계를 간선으로 할당한다.
단계 2. 하나의 정점에 대해서 주변의 정점들을 클러스터로 그룹핑한다.
단계 3. 클러스터를 새로운 정점으로 할당하고 클러스터들 사이의 새로운 간선을 생성한다.
단계 4. 정점과 간선을 가지고 있는 새로운 그래프를 정의한다.
단계 5. 정점의 수|V| 와 최종 추상화의 크기를 결정하는 계수값을 비교한다. 조건에 부합하면 작업을 종료하고, 그렇지 않으면 단계 2로 되돌아간다.

그러므로 각 레벨에서 정의되는 중간 그래프 G_i 는 그 자체로서 의미를 가지게 된다. 왜냐하면 정점은 단백질 그룹을 의미하고 간선은 비슷한 기능을 하는 정점들 간의 연결이기 때문이다. 그림 2는 제안하는 기능 기반 추상화 방법의 예를 보여준다. 원은 단백질을 나타내며 그 사이 간선은 단백질사이의 상호관계, f1,f2 등은 단백질이 가지고 있는 기능이다. 먼저 추상화는 그림2-a의 검게 표시된 정점으로부터 시작된다. 그 이유는 앞서 설명한대로 추상화의 시작시점은 $|\Gamma(v, F)|$ 값을 따르기 때문이다. 각각의 클러스터는 그림에서 점선으로 표시한 것처럼 좌측부분은 f1, 가운데는 f2, 우측은 f3의 기능으로 추상화되어 그림 2-b처럼 3개의 정점을 갖는 그래프로 다시 생성된다. 이렇게 구성된 새로운 그래프는 미리 정의된 최종 추상화의 크기를 결정하는 계수값에 의해서 다시 추상화를 하게 될지 결정된다. 이 경우 계수 값을 2로 주었다고 가정하면, 그림 2-b의 점선 부분과 같이 기능 f1을 중심으로 다시 추상화를 하게 되며 최종적으로 그림 2-c와 같은 정점 2개만을 가지는 그래프 레이아웃을 생성하게 된다. 그림 2-a와 2-c를 비교하게 되면 다음과 같은 사실을 알 수 있다. 실제 9개의 정점으로 구성된 원래 그래프가 두개의 정점을 갖는 단백질 a, 단백질 b로 구성되고 각각 (f1,f2), (f3,f4)의 기능을 갖는 것으로 추상화 되었다. 이것은 생물학자들에게 이 그래프를 통해서 단백질 상호 작용뿐만 아니라 단백질 기능의 관계도 파악할 수 있게 해준다.

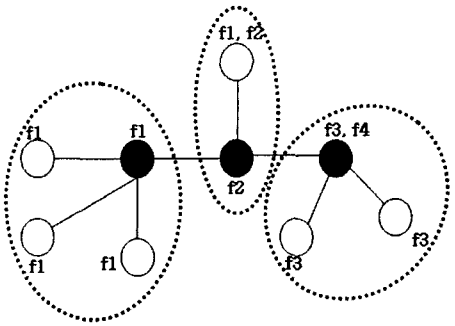

그림 2-a. 오리지널 그래프

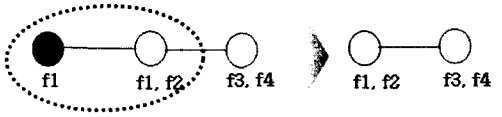

그림 2-b. 추상화된 그래프 (레벨1)

그림 2-c. 추상화된 그래프 (레벨2)

2.2. 위치화

위치화 부분에서는 그래프를 그리기 위해 각각의 정점들의 위치가 결정되어 진다. 위치화를 위한 알고리즘으로는 FDP (force-directed placement)[6] 혹은 spring-embedded 알고리즘[7] 등이 있는데 제안하는 방법에서는 FDP 알고리즘을 이용한다.

2.3. 확장

확장 부분에서는 최종 그래프로 추상화 되었다가 다시 각각의 중간 그래프로의 확장에 관한 내용을 담고 있다. 만일 우리가 새로 정의된 중간 그래프 $G_i=<V_i, E_i>$를 가지고 있다면 이 그래프는 그 부모 그래프인 $G_{i-1}=<V_{i-1}, E_{i-1}>$로 확장된다. 예를 들어 부모레벨의 매칭된 한 쌍의 정점을 v1,v2라 하면 v1,v2 $\in V_{i-1}$ 이다. 이때 두 정점이 추상화 된다고 하면 현 레벨에서는 v1,v2가 하나의 정점으로 표현되고 이것을 v라고 하면 $v \in V_i$ 로 표현된다. 추상화의 반대개념인 확장에서는 그림4 처럼 G_i가 G_{i-1}로 확장될 때의 정점들의 위치를 나타내는 역 추상화에 대한 설명이다. 기능 f1과 f2를 가지고 있는 정점이 확장될 때 스프링 상수 k(반지름 값)를 이용하여 원 둘레로 확장된다. 원주 상에 나타날 정점의 개수에 따라서 다음의 각도로 배치된다.

$$\text{정점들사이의각도} = \frac{360}{\text{원주상에나타날정점의개수}}$$

즉 그림 4의 경우는 f1,f2의 기능을 가진 정점이 f1,f2와 f1의 기능을 가진 정점으로 확장된다. 즉 원주 상에 나타날 정점의 개수는 2개 이며 둘 사이의 각도는 180도의 위치로 확장된다.

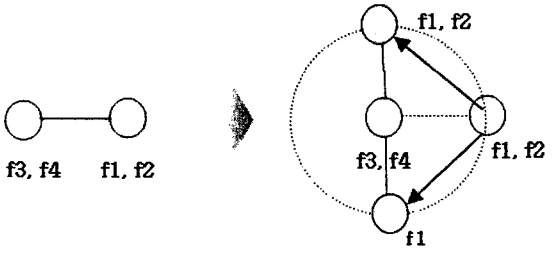

그림 4. 확장

3. 결론

이 논문에서는 단백질의 기능파악과 관련하여 기존에 알려져 있는 단백질의 상호 관계를 시각화 하는 방법에 대하여 제안하였다. 단백질 데이터가 대용량이라는 이유 때문에 시각화를 하여도 제대로 파악할 수 없다는 단점을 가지고 있다. 제안하는 방법은 단백질이 가지고 있는 기능과 관련하여 단백질 데이터를 추상화 하는 방법으로서 기존의 그래프 레이아웃 방법에 비해 생물학자들이 쉽게 단백질 기능과 그 상호작용을 파악할 수 있도록 훨씬 간결한 그래프를 제공한다. 또한 사용자들은 단백질의 기능을 고려해서 구성된 추상화 방법으로 레벨별 중간 그래프를 볼 수 있으며 각각의 중간 그래프는 단백질을 분석하고 상호작용을 파악하는데 도움을 주기 때문에 기존의 방법에 비해서 훨씬 많은 정보를 제공해 줄 것으로 기대된다.

참고문헌

[1] J. S. Richardson & D. C. Richardson, "in Prediction of Protein Structure and the Principles of Protein Conformation" (G. D. Fasman, ed.), Plenum Press, N.Y. (1989) pp. 1-98.

[2] A. Vazquez, A. Flammini, A. Maritan and A. Vespignani, " Global protein function prediction in protein-protein interaction networks", arXiv: cond-mat /0306611 v1 24 Jun 2003

[3] Byong-Hyon Ju, Byungkyu Park, Jong H.Park and Kyungsook Han "Visualizaion and analysis of protein interactions" BIOINFORMATICS APPLICATIONS NOTE Vol. 19 no 2 2003 p 317-318

[4] Anton J. Enright and Christos A.Ouzounis "BioLayout-an automatic graph layout algorithm for similarity visualization" BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no 9 2001 p 853-854

[5] Peter Uetz, Tery Ideker and Benno Schwikowski. T,M,J. "Visualization and Integration of Protein-Protein Interactions". 1-28 (http://www.systemsbilogy.org/pubs/vizprotein)

[6] T,M,J. Fruchterman and E. M. Reingold. "Graph Drawing by Force-Directed Placement". Software Practice & Experience, 21(11):1129-1164, 1991

[7] P.Eades. "A Heuristic for Graph Drawing". Congressus Numerantium, 42:149-160, 1984