단백질 상호작용 네트워크에서 상동성 기반 바이오 콤플렉스 예측

최재훈+ 박종민 박수준
한국전자통신연구원
{jhchoi+,jmpark93,psj}@etri.re.kr

A Homology-Based Prediction of Biological Complexes in a Protein-Protein Interaction Network

Jae-Hun Choi+, Jong-Min Park, Soo-Jun Park
Electronics and Telecommunications Research Institute(ETRI)

요 약

본 논문에서는 생물학적 실험에 의해 추출된 이종의 단백질 콤플렉스를 통해 대상 종의 콤플렉스를 단백질 상호작용 네트워크에서 예측할 수 있는 방법을 제안한다. 이 예측은 먼저 이종 사이에 단백질들의 비교를 통해 상동성을 색인한 다음, 이 상동성을 이용하여 이종의 콤플렉스를 대상 종으로 변형하고 그 형태를 단백질 상호작용 네트워크에서 탐색하는 과정으로 수행된다. Swiss-Prot 데이터베이스의 단백질들을 대상으로 상동성 색인을 색인하였으며, 콤플렉스 형태를 분석하기 위해 DIP의 단백질 상호작용 네트워크를 이용하였다.

1. 서론

단백질 상호작용 네트워크(PPI Network)는 세포의 통합 시스템으로서 생체 메커니즘들을 전체적으로 이해하는데 중요한 데이터이다. 단백질은 생물학적으로 고유한 기능들을 가지며, 다른 단백질과의 유기적인 상호작용을 통해 다양한 생명현상의 주도적 역할을 수행한다. 이 단백질 상호작용에 의해 수행되는 역할들은 PPI 네트워크에서 콤플렉스로 존재한다. 따라서, PPI 네트워크에는 많은 콤플렉스들이 포함되어 있으며, 이들은 생체 운영의 기본적인 단위로서 간주된다. 대표적인 예로 'Hemoglobin', 'RNA Polymerase' 등이 있다. 특히, 질병과 밀접하게 관련된 특별한 콤플렉스들은 신약개발에 핵심적으로 이용되고 있다. 또한, 이 콤플렉스에 포함된 단백질 상호작용 관계의 분석을 통해 신약물질의 적절성을 예측할 수 있다.

현재, PPI 데이터들은 'Yeast Two-Hybrid' 와 같은 대용량 실험 방법을 통해 빠르게 생성되고 있다. 또한, 이들 중에 포함된 많은 단백질 콤플렉스들도 'TAP-MS'와 'HMS-PCI' 같은 실험에 의해 식별되고 있다. 그러나, 이 방법들은 사용자가 요구하는 구체적인 콤플렉스를 탐색하기 위해서는 반복되는 실험을 수행해야 하기 때문에 많은 비용이 요구된다.

이 이유로 실험 이전에 잘 알려진 콤플렉스를 분석하여 탐색하거나 새로운 콤플렉스를 예측할 수 있는 방법들이 개발되고 있다. 대표적으로 [1]에서는 PPI 네트워크에서 클릭(Clique) 형태의 밀집된 부분 네트워크를 콤플렉스로 예측하였다. 최근, Bader and Hogue[2]는 클릭 탐색 방법을 개선하여 비슷한 기능을 하는 단백질들이 상호작용을 하면서 동시에 밀집된 부분 네트워크를 탐색하기 위한 클러스터링 방법을 제안하였다. 이 방법은 세포 내에서 콤플렉스에 포함된 단백질들이 서로 유사한 기능을 하면서 긴밀하게 상호작용을 한다는 가정을 기반으로 하고 있다. 그러나, 이 크러스터링은 유사하지는 않지만 서로 연관된 기능을 수행하는 단백질들로 구성된 콤플렉스를 예측할 수 없다는 단점을 가지고 있다.

다른 방법으로 Leser[3]는 사용자가 요구하는 콤플렉스를 탐색할 수 있는 질의 언어 PQL(Pathway Query Language)를 정의하였다. SQL과 유사한 형태의 이 PQL은 네트워크에서 사용자가 요구하는 두 단백질들 사이의 경로를 표현할 수 있다. 특히, 'WHERE' 절에 다양한 제약 사항을 기술할 수 있기 때문에 상대적으로 다양한 구조의 콤플렉스를 탐색할 수 있다. 예를 들어, 단백질의 속성이나 두 단백질들 사이의 경로 조건 등을 제약사항으로 기술한다. 그러나, 이 방법은 사용자가 직접 콤플렉스 형태를 PQL로 기술해야 한다는 단점을 가지고 있다.

이 단점을 보완하기 위해 본 논문에서는 생물학적 실험에 의해 추출된 이종의 단백질 콤플렉스를 통해 대상 종의 콤플렉스를 단백질 상호작용 네트워크에서 예측할 수 있는 방법을 제안한다. 이 예측은 먼저 이종 사이에 단백질들의 비교를 통해 상동성을 색인한 다음, 이 상동성을 이용하여 이종의 콤플렉스를 대상

종으로 변형하고 그 형태를 단백질 상호작용 네트워크에서 탐색하는 과정으로 수행된다.

2. 단백질 콤플렉스 예측

생물학적으로 이종 사이에 단백질 및 이들의 상호작용에는 상당한 상동성이 존재한다. 따라서, 단백질 상호작용 관계로 표현되는 콤플렉스 역시 이종 사이에 상동성이 존재한다. 이 장에서는 이 상동성을 이용하여 이미 실험을 통해 밝혀진 이종의 콤플렉스로부터 대상 종의 콤플렉스를 예측할 수 있는 과정을 설명한다.

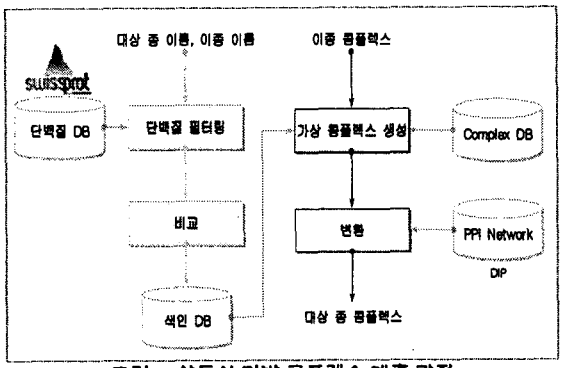

그림 1. 상동성 기반 콤플렉스 예측 과정

그림 1은 상동성 기반 콤플렉스 예측 과정을 설명하고 있다. 이 과정은 크게 단백질 색인 단계와 콤플렉스 변환 단계로 구분된다. 먼저, 단백질 색인 단계에서는 대상 종의 단백질과 유사한 생물학적 특성을 가지는 이종의 상동 단백질들을 색인한다. 다음으로, 이 색인 데이터를 이용하여 이종 콤플렉스들을 대상 종의 콤플렉스로 변환한다. 이때, Swiss-Prot 데이터베이스의 단백질들을 대상으로 상동성을 색인하였으며, 콤플렉스 형태를 분석하기 위해 DIP의 단백질 상호작용 네트워크를 이용하였다. 또한, 단백질 콤플렉스는 대부분 "yeast protein complex database"을 사용하였다.

3. 이종의 상동 단백질 색인

이종의 단백질 상호작용 관계로 표현된 콤플렉스를 대상 종의 콤플렉스로 변형하기 위해서는 해당 종에 대한 단백질의 상동성 관계를 파악해야 한다. 일반적으로 두 단백질들 사이의 상동성은 이들이 가지는 정보를 상호 비교함으로써 파악될 수 있다. 아미노산 서열은 단백질들 사이의 상동성을 파악하는데 매우 중요한 정보로 이용되며, BLAST와 같은 동적 프로그램(dynamic program)을 통해 그 유사도를 비교적 정확하게 평가할 수 있다[4]. 그러나, 하나의 종에 소속된 단백질의 수가 매우 많기 때문에 두 종 사이에 존재하는 모든 단백질들을 상호 비교하는 것은 많은 시간을 요구한다. 특히, 하나의 단백질에 대한 이종의 단백질들의 상동성 차이는 매우 극명하다. 따라서, 단백질 상동성 색인에 대한 성능 향상을 위해 서열 비교 이전에 이미 밝혀진 다른 정보를 이용한 필터링 단계가 반드시 요구된다.

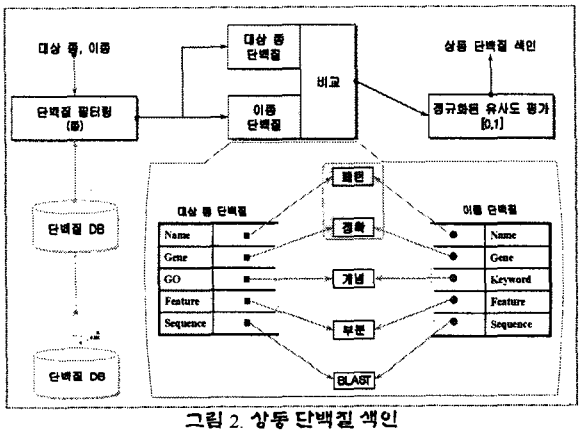

그림 2. 상동 단백질 색인

그림 2는 하나의 단백질과 유사한 이종의 단백질들을 색인하는 과정을 설명하고 있다. 먼저, 대상 단백질과 이종 단백질들을 데이터베이스에서 필터링한 다음, 단백질 비교 과정을 통해 대상 단백질과 유사도가 높은 이종 단백질들을 색인한다.

필터링에서는 하나의 대상 단백질과 이종의 단백질들에 대한 이름 패턴, 유전자, 키워드 그리고 서열 특징의 유사도를 비교한다. 이름에 대한 패턴 비교는 단백질에 대한 명칭과 이 단백질에 대한 유전자 명칭에 대한 문자열 패턴의 유사성을 평가한다. 이때, 각각의 명칭에 대한 동의 명칭 역시 함께 포함하여 유사성을 평가한다. 키워드 비교는 단백질의 특성을 명시하기 위해 사용한 GO(Gene Ontology) 용어들에 대한 개념적 거리를 통해 평가된다. 서열에 대한 특징은 부분 비교를 통해 많은 특징을 공유하는 두 단백질의 유사도를 높게 평가하였다. 이 필터링을 통해 높은 유사도를 가지는 단백질만을 후보로 선정한다.

필터링된 단백질들은 대상 단백질과 Needleman & Wunsch 의 전역 정렬 알고리즘을 통해 서열이 비교된다. 이때, 각각의 아미노산들 사이의 유사도 계산을 위해 유사 행렬 BLOSUM80 (Entropy=0.9868,Expected=-0.7442)이 사용된다. 서열 비교 유사도를 0과 1사이로 정규화하기 위해 대상 단백질 자신과 먼저 서열 비교를 수행한 다음, 그 유사도와 각각의 이종 단백질 서열

유사도의 비율을 계산한다. 따라서, 단백질 이름 패턴, 유전자, 키워드, 특징 그리고 서열의 각각의 유사도에 대해 일정한 임계값 이상인 단백질들에 대해서만 상동성 색인을 수행한다.

4. 상동성 기반 콤플렉스 변환

이 절에서는 상동 단백질 색인 데이터를 통해 이종 콤플렉스를 통해 대상 종의 콤플렉스를 예측하기 위한 과정을 설명한다. 이 과정은 대상 종에 대한 가상 콤플렉스를 생성하는 단계와 이 가상 콤플렉스를 예측 콤플렉스로 변환하는 단계로 구분된다.

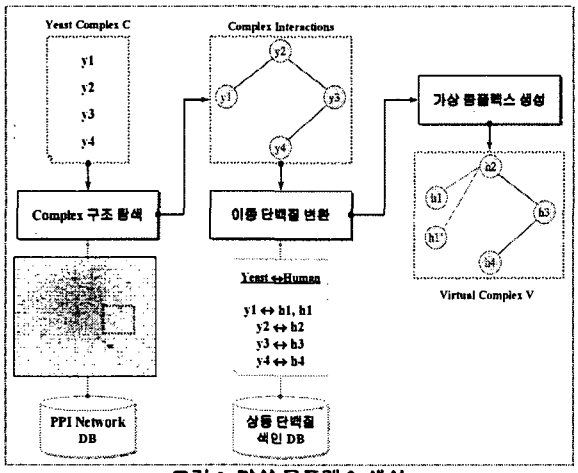

그림 3. 가상 콤플렉스 생성

[그림 3]은 가상 콤플렉스를 생성하는 단계를 설명하고 있다. 먼저, 이종의 PPI 네트워크에서 콤플렉스를 구성하는 단백질들의 PPI 관계를 탐색한다. 이 탐색된 이종의 콤플렉스의 PPI 관계들을 상동 단백질 색인 DB를 통해 대상 종의 상의 PPI 관계로 변환하여 가상의 콤플렉스를 생성한다. 이때, 이종의 하나의 단백질이 대상 종의 여러 단백질들로 변환될 수 있다. 예를 들어, 콤플렉스 C를 구성하는 단백질들(y1,..,y4)의 관계는 Yeast에 대한 PPI Network로부터 파악된다. 또한, y1는 대상 종인 Human의 두 개의 단백질 h1, h1 과 상동관계를 가지게 되어 C의 관계 <y1, y2>가 가상의 콤플렉스 V에서 <h1, h2> 그리고 <h1 ,h2 >으로 생성된다.

[그림 4]는 가상 콤플렉스를 예측 콤플렉스로 변환하는 과정을 설명하고 있다. 먼저, 가상 콤플렉스 V를 구성하는 단백질들이 실제 Human의 PPI 네트워크에서 어떠한 상호작용 관계를 이루고 있는지 탐색한다. 다음으로 가상의 콤플렉스에서의 PPI 관계와 탐색된 PPI 관계를 상호 비교하여 대상 종에 대한 콤플렉스를 예측한다. 예를 들어, 예측된 콤플렉스 P에 존재하는 관계 <h1, h1 >은 가상 콤플렉스 V에는 존재하지 않기 때문에 탐색된 콤플렉스 D에서 유추된다. P의 단백질 h4와 관계 <h3, h4>는 반대로 V에서 유추될 수 있다.

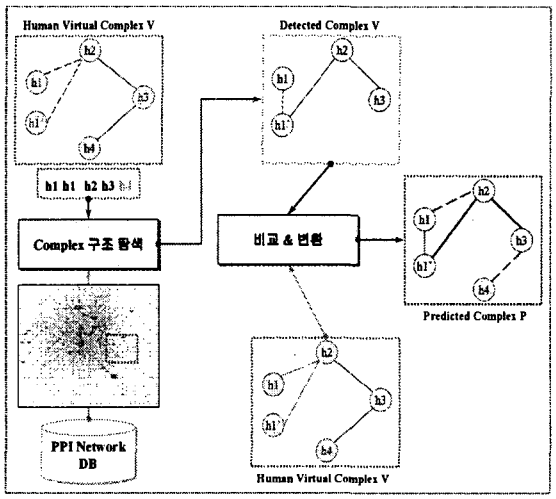

그림 4. 예측 콤플렉스 변환

5. 결론

본 논문에서는 이종들 사이에 존재하는 단백질 상호작용 관계의 상동성을 이용하여 특정 종의 콤플렉스를 예측할 수 있는 방법을 제안하였다. 이 방법을 통해 비교적 쉽게 생물학적 실험으로 추출되는 모델 생물의 단백질 콤플렉스를 통해 산업적 가치가 높은 인간과 같은 고등 생물의 콤플렉스를 예측할 수 있다. 향후에는 다양한 실제 실험 데이터를 적용하여 예측의 신뢰성을 향상시킬 필요가 있다.

참고문헌

[1] A.C. Gavin, M. Bosche, et al., "Functional organization of the yeast proteome by systematic analysis of protein complexes," Nature, 15(6868). 2002.

[2] G. D. Bader and C. W. Hogue, "An automated method for finding molecular complexes in large protein interaction networks," BMC Bioinformatics., 4:2, 2003.

[3] U. Leser, "A query language for biological networks," Bioinformatics, Vol.21 Suppl. 2:ii33-ii39, 2005.

[4] T.W. Huang, "POINT: a database for the prediction of ppi based orthologous interactome," Bioinformatics, Vol. 20, No. 17, 2004.