단백질 상호작용 관계의 상동성 기반 검증

최재훈+ 박종민 박선희
한국전자통신연구원
{jhchoi+,jmpark93,psh}@etri.re.kr

A Homology-Based Verification of Protein Interaction Relationships

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

요 약

본 논문에서는 생물학적 실험에 의해 추출된 특정 종의 단백질 상호작용 관계를 다른 여러 종에서 이미 밝혀진 단백질 상호작용 관계들을 통해 검증할 수 있는 방법을 제안한다. 이 검증을 위해 기본적으로 요구되는 이종간 단백질들 사이의 상동성 관계는 Swiss Prot 데이터베이스의 모든 단백질들에 대해 이름 패턴, 키워드, 서열 비교를 통해 구축된다. 즉, 특정 종에 대한 단백질 상호작용 관계를 여러 종의 단백질 상호작용 관계들로 상동화하고, 이 상동화된 관계들이 각각의 종에 어떠한 형태로 존재하는지의 여부를 판단함으로써 검증된다.

1. 서론

단백질은 유전자가 발현되어 생성되는 물질로서 생체 내에서 고유한 기능을 가지며, 다른 단백질과의 유기적인 상호작용을 통해 다양한 생명현상에 주도적 역할을 수행한다. 대표적으로, 생체 신호를 세포 핵까지 전달하여 생물학적 현상을 발현하는 신호전달, 세포의 생명 주기 및 발달, 물질에 대한 대사 등은 여러 단백질들의 복잡한 상호작용을 통해 수행된다. 따라서, 현대의 생명과학은 개개의 유전자나 단백질보다 이들 사이의 복잡한 상호작용을 통해 전체적인 관점에서 생명 현상을 규명하려는데 초점을 맞추고 있다.

단백질 상호작용은 생체 내에서 특정한 생물학적 작용이 수행되기 위해 여러 단백질들이 상호간에 형성하는 관계로 정의할 수 있다. 즉, 단백질 상호작용 관계는 하나의 단백질이 다른 단백질과 특정한 상호작용을 형성한다고 해석할 수 있다. 일반적으로 단백질 상호작용 관계는 이스트 투 하이브리드(yeast two hybrid)와 같은 대용량 방법(high-throughput screening)에 의해 실험되고 있다[1].

이 방법에서 두 단백질의 유전자는 변형되어 각각 베이트 단백질(bait protein)과 프레이 단백질(prey protein)로 발현된다. 즉, 베이트 단백질(bait protein)은 보고 유전자의 프로모터와 결합할 수 있는 DNA 결합 부위(DNA binding domain)을 갖게 된다. 그리고, 프레이 단백질(prey protein)은 보고 유전자를 발현시킬 수 있는 전사 활성화 부위(transcription activating domain)를 갖게 된다. 이때, 보고 유전자가 발현된다면 베이트 단백질과 베이트 단백질의 특정 도메인이 서로 상호작용을 한다고 말할 수 있다. 이 실험을 통해 구축된 데이터베이스로 PIM, BIND, DIP, GRID 등이 있다.

일반적으로 하나의 종에서 매우 방대한 단백질 상호작용 관계가 추출되고 있다. 예를 들어, DIP에 의하면 사카로미세스 세레비시아(Saccharomyces Cerevisiae)는 4,772 단백질 그리고 15,479 상호작용 관계를 가지고 있다고 알려져 있다. 그러나, 이들 중에는 실제로는 상호작용을 하지 않는 많은 오류(false positive)를 포함하고 있다. 이 오류를 검출하기 위해서 면역침강(co-immunoprecipitation)과 같은 생물학적 실험을 수행할 수 있으나, 방대한 단백질 상호작용 관계에 대해 이 실험을 수행하기에는 매우 많은 비용이 요구된다.

현재, 많은 연구들이 단백질 상호작용 검증 보다는 예측에 집중되어 진행되고 있다. 이 예측은 크게 기계 학습 방법[2]과 단백질 상동성 방법[3]으로 구분된다. 그러나, 이들 방법 역시 다음에 기술되는 이유로 많은 오류(false positive)를 가지고 있다. 따라서, 단백질 상호작용 관계에 대한 검증 방법이 관계 데이터의 신뢰성 확보를 위해 반드시 요구되고 있다.

전자 예측에서는 단백질을 특성(서열, 도메인, 발현 등)으로 기술하기 때문에 기존에 실험에 의해 밝혀진 많은 단백질 상호작용 관계를 특성들의 관계 데이터로 표현할 수 있다. 이 데이터에 대한 기계학습을 통해 특성들 사이의 관계 규칙을 추출하고, 이 규칙을 통해 새로운 단백질 상호작용 관계를 예측하게 된다. 그러나, 이 방법은 규칙의 범위도를 높이면 신뢰도가 현격하게 떨어지는 오류(false positive)와 신뢰도를 높이면 범위도가 현격하게 떨어지는 오류(false negative)를 가지고 있다.

후자 예측은 이종간 단백질 상동성 관계와 단백질 상호작용 관계를 통해 특정 종의 단백질 상호작용 관계를 예측한다. 즉, 이종에 두 단백질들 사이에 상호작용 관계가 존재한다면, 해당 종의 상동성 단백질들 사이에도 상호작용 관계가 존재할 것으로 예측한다. 이때, 예측을 위해 사용되는 단백질 상호작용 관계가 많은 오류(false positive)를 가지고 있다면 예측되는 단백질 상호작용 역시 같은 오류(false positive)를 가지게 될 것이다. 또한, 이 방법은 특정 몇 개 종에서만 존재하는 단백질 상호작용 관계를 예측할 수 없는 오류 (false negative)를 가지고 있다.

본 논문에서는 생물학적 실험에 의해 추출된 특정 종의 단백질 상호작용 관계를 다른 여러 종에서 이미 밝혀진 단백질 상호작용 관계들을 통해 검증할 수 있는 방법을 제안한다. 이 검증을 위해 기본적으로 요구되는 이종간 단백질들 사이의 상동성 관계는 Swiss Prot 데이터베이스의 모든 단백질들에 대해 이름 패턴, 키워드, 서열 비교를 통해 구축된다. 즉, 특정 종에 대한 단백질 상호작용 관계를 여러 종의 단백질 상호작용 관계들로 상동화하고, 이 상동화된 관계들이 각각의 종에 어떠한 형태로 존재하는지의 여부를 판단함으로써 검증을 수행하게 된다. 따라서, 이 검증은 생물학적 실험에 포함된 많은 오류(false positive)를 보정할 수 있게 한다.

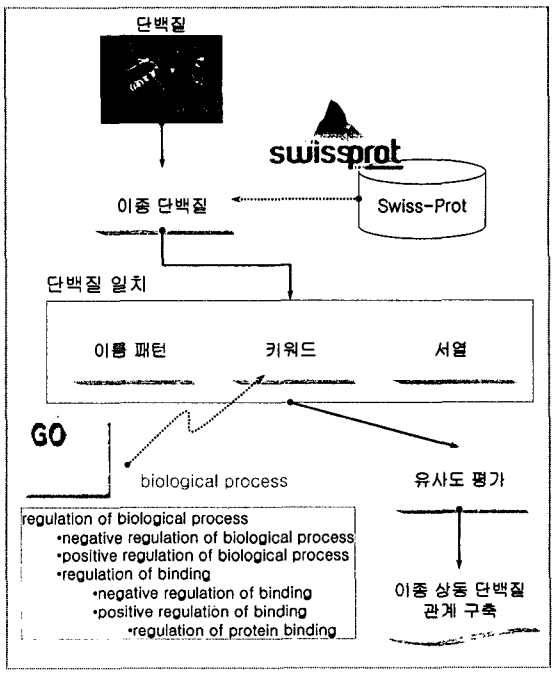
2. 단백질 상동성 관계

특정 종에 존재하는 단백질 상호작용 관계를 검증하기 위해 다른 종에는 이와 상동성이 있는 관계가 어떤 형태로 존재하는지를 파악할 수 있어야 한다. 하나의 단백질 상호작용 관계에 대한 이종간 상동성을 파악하기 위해서는 먼저 이 관계 포함된 단백질에 대한 상동성을 파악해야 한다. 두 단백질 상동성 비교는 BLAST와 같은 아미노산의 서열 정렬(Sequence Alignment)을 통해 수행된다. 그러나, 현재 공개되어 있는 단백질 데이터베이스를 이용한다면 아미노산 서열 외에도 상동성을 비교할 수 있는 다양한 특성들을 제공한다.

본 논문에서는 Siwss-Prot 데이터베이스의 단백질에 대해 이름 패턴, 키워드 그리고 서열 일치 방법을 사용한다. 이름에 대한 패턴 일치에서는 단백질에 대한 명칭과 이 단백질에 대한 유전자 명칭에 대한 스트링 패턴의 유사성을 평가한다. 이때, 각각의 명칭에 대한 동의 명칭 역시 함께 포함하여 유사성을 평가한다. 키워드 일치는 Swiss Prot에서 단백질의 특성을 명시하기 위해 사용한 용어들에 대한 개념기반 유사성으로 평가된다. 두 키워드들 사이의 개념기반 유사성을 위해 온톨로지(GO:Gene Ontology)를 이용한다. 즉, 온톨로지에서 두 용어 사이의 개념적 거리를 통해 유사성을 평가할 수 있다. 서열 일치는 BLAST에서 사용하는 서열에 대한 지역 정렬 방법을 사용하였으며, 서열에 대한 특성 정보 역시 유사성 평가에 이용된다. 단백질 서열에 대한 특성 정보는 Swiss-Prot에서 상세하게 제공하고 있다.

[그림 1]은 이종간 단백질 상동 관계를 구축하기 위한 과정을 설명하고 있다. 먼저, 하나의 단백질과 Swiss-Prot의 모든 이종 단백질을 일치시킨 다음 일정한 유사도의 단백질에 대해 상동 관계를 설정하게 된다. 유사도 평가는 위에서 기술한 3가지 일치 방법에 의해 구해진 각각의 유사도에 일정한 가중치를 도메인 특성에 따라 부여할 수 있다. 즉, 서열 일치와 같은 중요한 일치 방법에 높은 가중치를 부여할 수 있다.

이름 패턴에서는 단백질 명칭 또는 동의 명칭을 용어 토큰으로 분리한다. 이 토큰의 일치 개수와 단백질에 대한 유전자 명칭의 일치에 따라 유사도가 결정된다. 즉, 유전자 명칭이 일치되고 단백질 명칭에 대해 많은 토큰이 일치될수록 이종 단백질의 유사도가 높게 평가된다. 토큰 일치 정도는 분리된 총 토큰 그리고 공통으로 나타나는 토큰의 개수가 각각 많을수록 높게 계산된다.



[그림 1] 단백질 상동 관계 구축

키워드에 대한 유사도는 두 단백질에 명시된 용어들의 개념기반 일치에 따라 평가된다. 이때, 총 용어의 개수와 유사한 용어의 개수가 각각 많을수록 높은 유사도를 가지게 된다. 두 용어 사이의 개념기반 유사도는 $e^{-0.3\times DIST(o_1,o_2)}$에 의해 평가된다. 즉, GO에서 ' regulation of binding' 과 ' regulation of protein binding' 은 개념 거리가 3이기 때문에 $e^{-0.3\times 3}$로 계산된다.

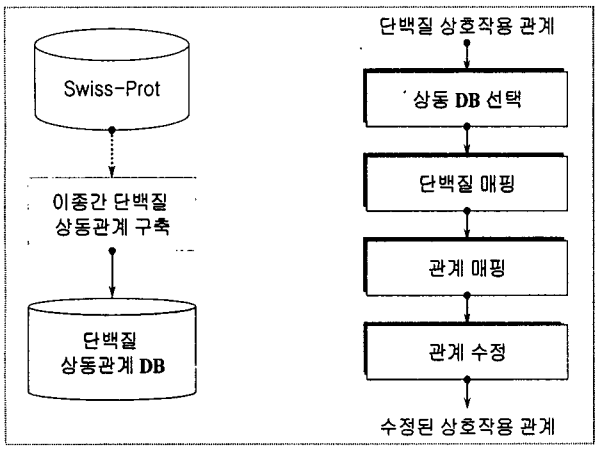
서열 유사도는 BLAST에 의한 서열 일치 정도와 데이터베이스에 기술된 서열 특징에 대한 일치 정도에 의해 계산된다. BLAST에 의한 일치 정도를 정규화하기 위해 자기 자신과 먼저 서열을 비교한다. 특징에 대한 일치 정도는 총 특징의 개수와 공통된 특징의 개수에 따라 계산된다.

이 3개의 요소에 의해 계산된 유사 정도들에 대해 각각의 가중치를 부여하여 최종 유사도를 계산한다. 즉, 도메인에 따라 특정 부분에 높은 가중치를 부여하게 된다. 이때, 최종 유사도가 일정한 값(threshold) 이상인 단백질과 상동관계를 형성한다. 즉, 하나의 단백질과 이종 상동 관계를 가지는 단백질은 여러 개가 존재할 수 있다.

3. 단백질 상호작용 관계 검증

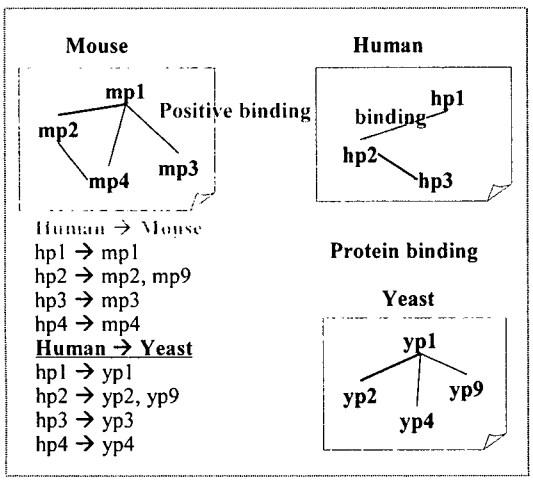
모든 종에 대해 이종간 단백질 상동 관계가 데이터베이스로 구

축되면 특정 종의 단백질 상호작용 관계에 대한 검증을 수행할 수 있다. 이 검증은 먼저 데이터베이스에서 상동 종 선택, 단백질 매핑, 관계 매핑, 그리고 관계 수정의 절차로 수행된다. [그림 2]는 이 과정을 설명하고 있다.



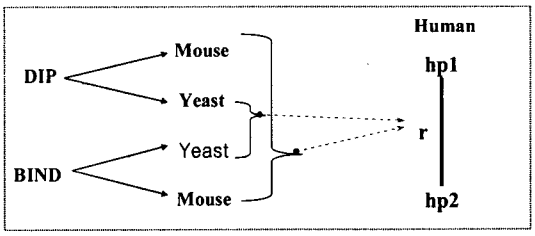
[그림 2] 단백질 상호작용 관계 검증

먼저 검증을 하고자 하는 목표 종의 단백질 상호작용 관계에 대해 상동 관계에 있는 대상 종을 선택한다. 따라서, 목표 종의 단백질 상호작용 관계의 단백질과 관계를 여러 상동의 대상 종에 존재하는 단백질과 관계로 매핑할 수 있다. 이 매핑을 통해 목표 종의 관계와 대상 종의 관계들을 비교함으로써 목표 종의 단백질 상호작용 관계가 적절한지를 검증한다.



[그림 3] 단백질 상호작용 관계 검증 예

[그림 3]은 이 검증 과정을 예를 통해 설명하고 있다. 먼저, 목표 종 Human에 대해 관계 (binding, hp1, hp2)의 단백질 hp1과 hp2는 대상 종 Mouse의 mp1와 mp2 그리고 대상 종 Yeast에 대해 yp1과 yp2로 각각 매핑된다. 따라서, 목표 종의 관계는 Mouse와 Yeast 관계 (positive binding, mp1, mp2)와 (protein binding, yp1, yp2)으로 검증될 수 있다. 즉, 대상 종에서 관계 종류에서 protein binding이 positive binding과 binding보다는 하위 개념이기 때문에 (binding, hp1, hp2)은 (protein binding, hp1, hp2)로 검증될 수 있다. 이 검증은 도메인에 따라 변형될 수 있다. 예를 들어, 구체적인 관계 정보보다 일반적인 정보를 원할 경우에는 (binding, hp1, hp2) 자체가 Mouse와 Yeast 데이터로부터 검증되었다고 말할 수 있다. 또한 [그림 4]와 같이 여러 단백질 상호작용 관계 데이터베이스를 이용할 수 있다. 즉, DIP과 BIND 데이터베이스에 대해 Mouse와 Yeast의 단백질 상호작용 관계를 이용하여 같은 방법으로 Human의 단백질 상호작용 관계를 검증할 수도 있다. 이런 검증은 단일 데이터베이스를 이용하는 검증보다 높은 신뢰도를 보장할 수 있을 것이다.



[그림 4] 여러 데이터베이스를 이용한 검증

4. 결론

본 논문에서는 특정 종의 단백질 상호작용 관계를 검증할 수 있는 방법을 제안하였다. 이 방법은 특정 종에 대한 단백질 상호작용 관계를 여러 종의 단백질 상호작용 관계들로 상동화하고, 이 상동화된 관계들이 각각의 종에 어떠한 형태로 존재하는지의 여부를 판단함으로써 검증을 수행하게 된다. 단백질 상호작용 관계의 상동성을 판별하기 위해 이종간 단백질들 사이의 상동성 관계를 이름 패턴, 키워드, 그리고 서열 비교를 통해 구축하였다. 이 상동 관계에 의한 단백질 상호작용 관계 검증은 생물학적 실험에 포함된 많은 오류(false positive)를 적은 비용으로 상당부분 보정할 수 있게 한다는 장점을 가지고 있다.

참고문헌

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

[2] S. M. Gomez, W. S. Noble, and A. Rzhetsky, "Learning to Predict Protein-Protein Interactions from Protein Sequences", Bioinformatics, Vol. 19, No. 15, 2003.

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