

단백질 상호작용 네트워크를 위한 템플릿 기반 동적 관리

박종민⁰ 최재훈 박선희
한국전자통신연구원
{ jmpark93⁰, jhchoi, shp }@etri.re.kr

A Template-based Dynamic Management of PPI Networks

JongMin Park⁰ JaeHun Choi SeonHee Park
Electronics & Telecommunications Research Institute(ETRI)

요 약

본 논문에서는 단백질 상호작용 네트워크를 효율적으로 구축하고 관리하기 위한 템플릿 기반 동적 관리 방법에 대해 제안한다. 기존의 단백질 상호작용 네트워크를 지원하는 대부분 시스템들은 각각의 단백질을 참조 데이터베이스에 대한 레퍼런스로 표현하고 있다. 따라서, 단백질에 상세한 정보를 가지고 분석하기 위해서는 참조 데이터베이스를 통해 상세한 정보를 얻어야 가능하다. 하지만, 방대하고 복잡한 상호작용 네트워크인 경우 분석에 많이 시간이 필요하며, 참조 데이터베이스의 레퍼런스가 변경된 경우 제대로 된 분석 결과를 얻을 수 없다. 본 논문에서 제시한 템플릿 기반의 동적 관리 기법은 다양한 사용자들이 자신의 요구에 맞는 단백질 상호작용 네트워크를 템플릿 검색을 통해 손쉽게 구축할 수 있도록 지원하며, 다중 사용자들이 서로간의 간섭 없이 각자의 단백질 상호작용 네트워크 관리할 수 있다. 또한, 참조 데이터베이스의 수정된 내용이 단백질 상호작용 네트워크에 지속적으로 반영할 수 있도록 한다.

1. 서 론

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[1]. 이러한 단백질들 사이의 상호작용 관계들을 단백질은 노드로, 이들 사이의 상호작용을 링크로 표현하면 네트워크로 나타낼 수 있다[2]. 현재, 단백질 상호작용 네트워크는 보통 “Yeast Two-Hybrid” 라는 생물학적 실험을 통해 빠르게 추출 되고 있으며, 추출된 상호작용 네트워크는 BIND(Biological Interaction Network Database), DIP(Database of Interacting Protein) 등과 같이 데이터베이스에 체계적으로 관리되고 있다.

단백질 상호작용 네트워크는 방대한 단백질들 사이의 복잡한 관계들로 표현되며 생물학자가 특정 단백질이 아닌 전체적인 생체 메커니즘을 이해할 수 있도록 한다. 그러나, 사용자들이 방대하고 복잡한 상호작용 네트워크를 직접 관리하고 분석할 수 없기 때문에 이를 지원하는 응용 시스템이 요구되고 있다. 단백질 상호작용 네트워크를 지원하는 응용 시스템에서 분석 및 관리를

효율적으로 수행하기 위해서는 상호작용 네트워크에 참여하는 단백질에 대한 상세한 정보가 필요하고, 이미 밝혀진 단백질에 대한 정보는 공유 가능해야 하며, 여러 사용자들이 서로의 간섭 없이 자신의 요구에 맞는 네트워크를 구축 및 관리 할 수 있어야 한다.

2. 관련연구

단백질 상호작용 네트워크를 지원하는 시스템으로는 InterViewer, Osprey, Cytoscape 등이 있으며, 대부분의 시스템에서 상호작용 네트워크에 참여하는 단백질을 표현하기 위해 사용되는 참조 데이터베이스로 ‘Swiss-Prot’이 많이 사용된다. Swiss-Prot은 단백질에 대한 분류(taxonomy), 유전자(gene), 인용된 문헌, 생물학적 기능, 서열 및 서열 특징 등에 대한 체계화된 정보를 유지하고 있다. 대부분의 시스템에서 사용하는 단백질 상호작용 네트워크 데이터는 단백질 노드에 대한 간단한 레퍼런스와 이들간의 상호작용 관계들만으로 작성되며, 상호작용 네트워크 데이터를 있는 그대로 읽고 물리적인 관계들을 이용한 분석만을 수행한다. 실제 단백질 상호작용 네트워크를 직접

구축하고 관리할 수 있는 시스템은 InterViewer, Osprey 가 있다.

InterViewer[3] 의 경우 단백질에 대한 상세한 정보 없이 참조 데이터베이스에 대한 레퍼런스로 상호작용 네트워크를 표현하며 분석 및 관리를 위한 기능들 대부분 물리적인 관계들을 사용하여 수행한다. 따라서, 단백질에 대한 상세한 정보를 필요로 하는 분석 작업 수행 시에 참조 데이터베이스를 통해 정보를 얻기 위한 번거로운 작업 수행하여야 한다. Osprey[4] 도 참조 데이터베이스에 대한 레퍼런스로 상호작용 네트워크 표현하지만 단백질에 대한 세부적인 정보까지 참조한다. 따라서, 분석 및 관리를 위한 기능들은 물리적인 상호작용뿐만 아니라 각각의 단백질들에 대한 세부적인 정보를 포함하여 처리한다. 하지만, 실험 및 분석 과정에 사용하기 위한 단백질을 직접 작성하거나 변경할 수 없으며, 기존에 작성된 상호작용 네트워크에 표현된 단백질의 레퍼런스가 참조 데이터베이스에서 변경된 경우 단백질에 대한 상세 정보를 얻을 수 없다.

따라서, 본 논문에서는 공개된 단백질에 대한 상세한 정보를 재사용 가능하도록 하며, 다중 사용자들이 자신의 의도에 맞게 새로운 단백질 생성하거나 수정하여 상호작용 네트워크를 구축할 수 있는 템플릿 기반 단백질 상호작용 네트워크 구축 방법에 대해 제안한다. 또한, 단백질 참조 데이터베이스는 주기적인 수정에 따라 단백질의 추가, 수정, 삭제 및 병합 등이 일어나므로 기존의 구축된 상호작용 네트워크의 단백질에 대한 데이터를 최신 데이터로 유지하기 위한 동적 관리 방법에 대해서도 설명한다.

먼저, 단백질 상호작용 네트워크의 분석 및 관리 과정에 필요로 하는 단백질에 대한 정보는 일부 제한적이고 대량의 단백질로 구성된 복잡한 상호작용 네트워크를 분석하는 작업은 많은 시간을 필요로 한다. 그러므로 본 논문에서 제안한 시스템에서는 Swiss-Prot 의 데이터들 중 단백질의 이름, 유전자, 분류, 생물학적 기능, 그리고 서열 및 서열 특징 등 필요한 정보만을 필터링하여 로컬 데이터베이스로 관리하고 부가적인 정보들은 직접적인 레퍼런스를 통해 참조 가능하도록 한다.

단백질 상호작용 네트워크를 관리 시스템에서는 기본적으로 사용자가 직접 새로운 단백질을 생성하고 수작업으로 단백질에 대한 정보를 작성할 수 있는 방법을 제공한다. 하지만, 기존의 밝혀진 단백질에 대한 정보들은 위에서 설명한 로컬 데이터베이스에 템플릿 기반 검색을 통해 얻을 수 있다. 템플릿 기반 검색에서는 일반적인 분류, 이름, 유전자 등에 대한 기본적인 불리언 검색뿐만 아니라 단백질의 생물학적 기능에 대한 GO(Gene Ontology)[5]를 이용한 개념 질의가 가능하다. 검색을 통해 얻어진 단백질은 사용자가 수정하여 사용할 수 있지만 작성된 단백질 상호작용 네트워크를 위한 별도의 데이터베이스로 관리되므로 다른 사용자에게 영향을 주지 않고 관리할 수 있다. 또한, 상호작용 관계에 대한 정의에서 관계의 종류도 UMLS(NLM's Unified Medical Language System)[6] 라는 체계화된 온톨로지를 사용하여 정의한다.

3. 템플릿 기반 단백질 상호작용 네트워크 구축

본 논문에서는 제안한 템플릿 기반 동적 관리 방법으로 단백질 상호작용 네트워크를 구축하는 과정을 그림 1에 도식화 하였다.

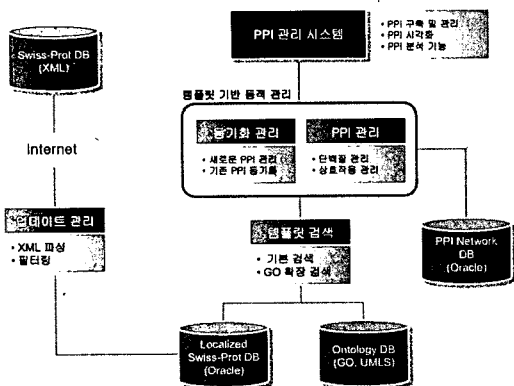


그림 1. 템플릿 기반 단백질 상호작용 네트워크 구축 과정

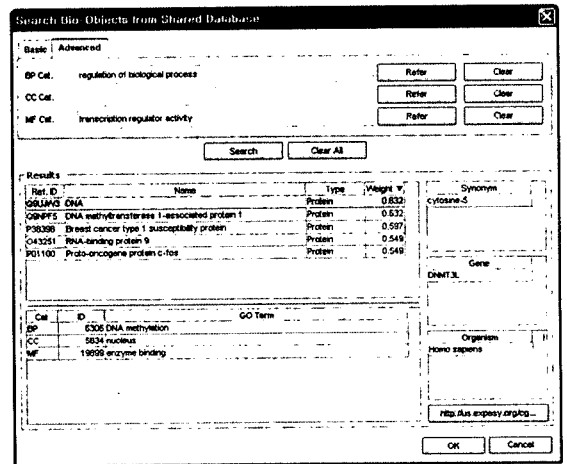


그림 2. 템플릿 검색을 위한 단백질 검색 다이얼로그

템플릿 검색의 예로서, 그림 2는 단백질의 분류가 'Homo Sapience' 이고, 생물학적인 기능(GO)은 BP(Biological Process)는 'regulation of biological process', MF(molecular function)는

'transcription regulator activity'인 개념에 직접적 또는 간접적으로 해당되는 기능을 가진 단백질을 검색하고 결과를 나타내고 있다.

4. 단백질 상호작용 네트워크를 위한 템플릿 기반 동적 관리

참조 레퍼런스뿐만 표현된 공개된 상호작용 네트워크나 본 논문에서 제안한 시스템으로 구축된 상호작용 네트워크의 단백질에 대한 최신 데이터를 계속적으로 유지하기 위해서는 자동화된 처리 방법이 필요하다. 또한, Swiss-Prot에는 "Secondary Acc No."라는 리스트 형태의 필드가 존재한다. 이 필드는 동일한 단백질로 판명되어 합쳐지거나 새로운 레퍼런스로 대체된 단백질에 대한 레퍼런스 리스트로 상호작용 네트워크가 작성된 시점에 따라 레퍼런스를 찾지 못하여 분석 과정에 참여시키지 못하는 문제가 있다.

본 논문에서 제안하는 템플릿 기반 동적 관리 방법은 사용자의 불필요한 간섭 없이 원하는 시점에 자동으로 단백질 상호작용 네트워크의 단백질 정보를 최신 데이터로 동기화하기 위한 방법이다. 사용자가 동기화를 수행하면, 기존 공개된 포맷으로 작성된 파일이나 데이터베이스의 단백질 상호작용 네트워크 데이터를 읽는 과정에서 참조 데이터베이스에 대한 레퍼런스를 사용하여 템플릿 검색을 통해 해당되는 단백질에 대한 정보를 가져와서 순차적으로 작성해 간다. 이때, "Secondary Acc No."에 해당되는 경우 "Primary Acc No."로 변환하여 검색하여 사용한다. 따라서, Swiss-Prot이 변경되더라도 사용자가 원하는 시점에 동기화 과정을 수행할 경우 위와 같은 과정을 반복적으로 수행하여 항상 최신의 단백질 상호작용 네트워크를 유지하도록 한다. 하지만, Swiss-Prot에서는 서로 다른 두 개 이상의 단백질이 동일한 단백질로 판명되어 합쳐지는 경우도 있다. 만약, 하나의 상호작용 네트워크에서 이런 경우가 발생하면 하나의 물리적인 단백질 노드를 만들고 상호작용 관계들을 병합한다. 또 다른 문제점은 사용자가 직접 단백질 데이터를 수정할 수 있으므로 동기화할 경우 사용자 정보와 변경된 단백질 정보를 처리하여야 한다. 모든 단백질에 대해 사용자가 수정한 내용을 기억하고 처리하기에는 성능 및 비용이 너무 많이 든다. 따라서, 사용자가 직접 수정한 단백질인 경우를 체크하여 리스트를 작성한 다음 기존 정보를 유지할 것인지, 해당 단백질에 대한 정보와 기존 정보를 병합을 수행할 지, 새로운 정보로 대체할지를 결정할 수 있도록 사용자 인터페이스를 제공한다. 그림 3은 본 논문에서 제안한 템플릿 기반 동적 관리 기법을 이용하여 작성된 단백질

상호작용 네트워크를 가시화한 그림이다.

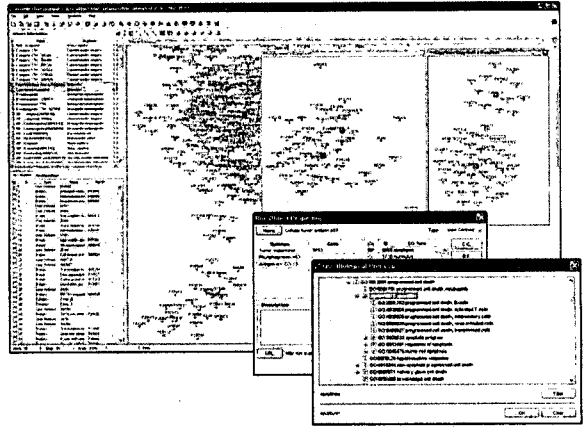


그림 3. 단백질 상호작용 네트워크 가시화

5. 결론 및 향후 연구 과제

본 논문에서 제시한 템플릿 기반의 동적 관리 기법은 다양한 사용자들이 자신의 요구에 맞는 단백질 상호작용 네트워크를 템플릿 검색을 통해 손쉽게 구축할 수 있도록 지원하며, 다중 사용자들이 서로간의 간섭 없이 각자의 단백질 상호작용 네트워크 관리할 수 있다. 또한, 참조 데이터베이스의 수정된 내용이 단백질 상호작용 네트워크에 대해 지속적으로 반영할 수 있도록 한다.

본 논문에서 제시한 템플릿 기반 동적 관리 기법은 단백질 상호작용 네트워크에 대해서만 적용된다. 하지만, 생물학적 상호작용 네트워크는 DNA, RNA, Gene 뿐만 아니라 단백질 복합체 (complex)를 포함하여 다양하다. 따라서, 본 논문에서 제안한 방법을 이러한 다양한 상호작용 네트워크에도 적용할 수 있는 유연한 방법에 대해 연구할 필요가 있다.

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] S. Oliver, "Guilt-by-Association Goes Global," Nature-News and Views, Vol. 403, 2000.
- [3] InterViewer, <http://interviewer.inha.ac.kr/>
- [4] Osprey, <http://biodata.mshri.on.ca/osprey>
- [5] Gene Ontology, <http://www.geneontology.org>
- [6] UMLS, <http://www.nlm.nih.gov/research/umls/>