PPI 네트워크에서의 래퍼 기반 단백질 식별

이용호° 최재훈 임명은 박수준
한국전자통신연구원
{louse°, jhchoi, melim, psj}@etri.re.kr

Wrapper-based Approach for Protein Identification in PPI Network

YongHo Lee°, JaeHun Choi, MyungEun Lim, SuJun Park
Electronics & Telecommunications Research Institute(ETRI)

요 약

단백질 상호작용 관계들은 고 성능 실험 기법을 이용한 생물학적 실험에 의해서 대규모로 추출되고, 동시에 이들을 구성하는 단백질 데이터 역시 공공 데이터베이스에 빈번하게 갱신되고 있다. 이 갱신으로 인하여 인터넷을 통해 공개되는 공공 데이터베이스와 PPI(Protein-Protein Interaction) 네트워크에 포함된 단백질 데이터가 서로 일치하지 않게 된다. 본 논문에서는 PPI 네트워크에 존재하는 단백질을 래퍼(Wrapper)를 이용하여 빈번하게 갱신되는 공공 데이터베이스의 단백질로 식별하고, 이 식별을 통해 PPI 네트워크에 존재하는 데이터들을 항상 최신 데이터로 동기화함으로써 데이터의 실시간성을 제공하고 데이터에 대한 신뢰도를 보장할 수 있도록 하였다.

1. 서 론

일반적으로 PPI(Protein-Protein Interaction) 네트워크는 단백질들 사이의 관계에 대한 집합으로 정의할 수 있으며, 세포의 생명 주기, DNA 복제 및 전사, 신호 전달, 물질 대사 등에 핵심적인 사항들을 명시하는 매우 중요한 데이터이다[1, 2]. 또한, 이들은 의료진단이나 신약개발과 같은 고부가가치 바이오 산업에 효과적으로 활용되고 있기 때문에 이들에 대한 중요도가 점점 증가하고 있다. 최근에 단백질 상호작용 관계들은 고 성능 실험 기법을 이용한 생물학적 실험을 통해 대규모로 추출되고 있다. 대표적인 데이터로 BIND(Biological Interaction Network Database), DIP(Database of Interacting Protein) 등이 있는데, 이런 데이터들은 주로 Swiss-Prot[3]이나 Gene Bank[4]와 같은 공공 데이터베이스로 유지 및 관리되고 있다. 또한, 이 데이터들을 계산 모델에 적용하여 예측된 결과들은 TrEMBL과 같은 공공 데이터베이스에 저장 및 관리되고 있으며, 이들 모두는 인터넷을 통해 공개되고 있다.

공개된 데이터베이스는 서로 다른 특성을 지니고 있는데, 그 특성은 보유 데이터 종류, 데이터 형태, 질의 처리 메커니즘, 시각적 질의 응답 여부로 특정 지어질 수 있다. 보유 데이터의 종류는 크게 특정 유기체 단백질들 간의 상호작용에 대한 정보만을 제공하는 데이터베이스와 일반적인 정보를 제공하는 데이터베이스로 분류된다. 질의 처리는 웹 기반 사용자 인터페이스의 제공 유무, 탐색어를 이용한 웹 기반 탐색 메커니즘의 제공유무, 그리고 그래픽 질의 결과의 제공 유무에 따라 특성화된다. 이런 이기종 데이터베이스에 대한 접근을 가능하게 하기 위한 이전의 연구에서는 데이터의 추출, 변형과 통합 과정에 래퍼 중재자(wrapper-mediator)[5, 6]를 사용하여 서로 다른 원천 데이터들에 대한 일관성 있는 접근 및 사용을 가능하게 하였다.

본 논문에서는 PPI 네트워크와 지역 데이터베이스 상호간의 데이터 일관성 유지 정책을 사용하고, PPI 네트워크에 존재하는 단백질을 래퍼(Wrapper)를 이용하여 빈번하게 갱신되는 공공 데이터베이스의 단백질로 식별하여 동기화함으로써 지역 데이터베이스, 전역 데이터베이스 그리고 PPI 네트워크를 구성한 단백질 데이터 상호간의 일관성을 유지하였다. 또한, 래퍼(Wrapper)로 데이터의 실시간성을 제공하여 PPI 네트워크에 존재하는 데이터들을 항상 최신 데이터로 유지함으로써 데이터에 대한 신뢰도를 보장할 수 있도록 하였다.

본 논문의 구성은 제 2장에서 관련 연구를 고찰하고 제 3장에서는 시스템 모델을 설명한다. 제 4장에서는 구현을 보이고 마지막으로 제 5장에서는 결론 및 향후 연구 과제에 대해 기술한다.

2. 관련 연구

생명과학 분야는 짧은 역사에도 불구하고 괄목할 만한 발전을 보이고 있으며, 다방면에 걸친 생명과학 연구들의 결과가 누적됨에 따라 유전자를 포함한 생명체에 관련된 정보량이 급속하게 증가하고 있는 추세에 있다. 생명과학 분야에서 생성되는 정보의 특징은 그 양이 막대할 뿐 아니라 정보의 관리가 서로 다른 조직이나 기관에 의해 수

행되고 있어 정보 분산성과 이산성의 정도가 심한 대용량 자료 환경을 구성하고 있다는 것이다. 이러한 생명과학 정보의 특성은 다양한 정보를 수집하고 조직화하여 빠른 기간 내에 비교 분석해야 하는 사용자들에게 심각한 문제점으로 대두되고 있다.

여러 기관에 의해 분산, 관리되는 정보들은 각기 다른 표현형태, 접근방식, 저장형태라는 측면에서 이질성을 보이고 있는데 이런 이질성을 극복하고 데이터의 추출, 변형과 통합 과정을 가능하게 하기 위한 방법으로 래퍼 중재자(wrapper-mediator)[5, 6] 기법이 사용되고 있다. 래퍼(Wrapper)는 형태와 저장 방법에 있어 이질성을 가지는 정보들을 일관된 방법으로 접근할 수 있는 기능을 제공하는데, 사용자 측면에서 이러한 기능은 이질적 원천 정보의 특성에 관계없이 다양한 정보에 대한 공통된 접근을 가능하게 하는 인터페이스로 볼 수 있다.

대표적인 이기종 데이터베이스에 대한 접근을 가능하게 하는 연구로는 TAMBIS[7], IBM DiscoveryLink[8]등이 있는데, 이 시스템들은 데이터의 추출, 변형과 통합 과정에 래퍼 중재자(wrapper-mediator) [5, 6]를 사용하여 서로 다른 원천 데이터들에 대한 일관성 있는 접근 및 사용이 가능하도록 하였다.

3. 시스템 모델

PPI(Protein-Protein Interaction) 네트워크와 지역 데이터베이스 상호간의 데이터 일관성을 위하여 지역 동기화 기법을 사용하고, PPI 네트워크와 전역 데이터베이스 상호간의 데이터 일관성을 위하여 전역 동기화 기법을 사용하였다. 이런 동기화를 통하여 정확한 단백질 식별이 가능하게 된다.

지역 데이터베이스는 Swiss-Prot[3] 데이터를 사용하여 구축되었고, 전역 동기화는 Swiss-Prot과 TrEMBL이 함께 있는 UniProt[9] 데이터를 이용하였다.

3.1. 래퍼(Wrapper)

이질적 원천 정보의 특성에 관계없이 원하는 내용만을 추출하기 위해서는 일관된 기술이 가능한 중간 언어가 필요하다. 이런 필요에 의하여 WDL이라는 중간 언어를 설계하였고, 래퍼(Wrapper) 변환 모듈을 개발하였다. 이렇게 함으로써, 래퍼(Wrapper)의 구조나 개발 언어에 대해서 전혀 모르는 사용자일지라도 추출하고자 하는 정보에 따라서 WDL만 작성하면 원하는 래퍼(Wrapper)를 개발할 수 있도록 하였다.

WDL(Wrapper Description Language)은 선정된 웹 사이트의 HTML 페이지에 따라서 추출하고자 하는 정보를 구조적으로 기술하는 스크립트 언어로서 래퍼(Wrapper) 개발을 위한 중간 언어이다. 웹 사이트로부터 원하는 정보만을 추출하기 위하여 자체적인 문법을 정의하였으며, Regular Expression을 사용하여 HTML 페이지로부터 원하는 텍스트의 추출을 더욱 쉽게 하였다.

사용자는 추출하고자 하는 정보가 있는 웹 사이트를 선정한 후, 원하는 요약 정보에 따라 WDL 파일을 작성한다. 그 후, 이 파일을 변환 모듈을 사용하여 JAVA 파일로 변환하고 래퍼(Wrapper) API를 이용하여 실행하게 된다. 래퍼(Wrapper)가 실행되면 HTML 페이지로부터 원하는 정보를 텍스트 형태로 추출하여 XML 형태로 변환하여 반환하게 된다. XML 형태로 반환하도록 한 이유는, 최근 생물학 분야에서 확장성과 편리함 때문에 실험 결과 및 데이터 관리에 XML형태의 데이터가 광범위하게 쓰이고 있기 때문이다. 그림 1 은 래퍼(Wrapper)의 작성 및 동작 순서를 보이고 있다.

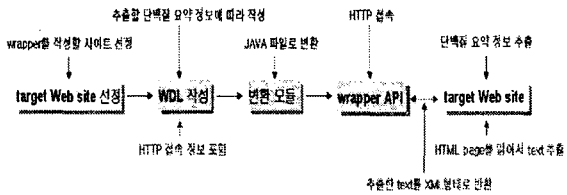

그림 1. 래퍼(Wrapper)의 작성 및 동작

3.2. 지역 동기화

지역 데이터베이스를 전역 데이터베이스의 데이터로 주기적으로 갱신하였을 경우 PPI 네트워크에서 사용자가 선택한 단백질의 데이터와 지역 데이터베이스의 데이터가 서로 다른 경우가 발생하는데, 이럴 경우 불일치를 해소하기 위하여 지역 데이터베이스의 데이터로 동기화하여 준다. 예를 들면, 단백질 P1은 서로 일치하는 데이터였지만 지역데이터베이스의 갱신으로 인하여 P1과 P1' 으로 서로 불일치가 발생한다. 이럴 경우 PPI 네트워크의 P1을 P1' 으로 동기화하여 준다. 그림 2 는 지역 동기화 과정을 보여주고 있다.

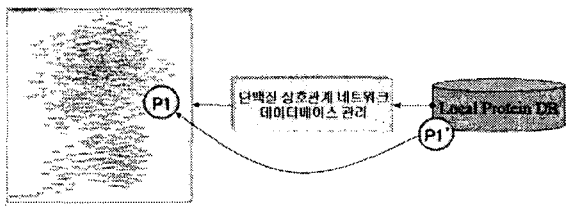

그림 2. 지역 동기화

3.3. 전역 동기화

생물학적 실험에 의하여 전역 데이터베이스에 있는 데이터가 변경되었을 경우 PPI 네트워크의 데이터와 서로 불일치가 발생할 수 있다. 이럴 경우 변경된 전역 데이터베이스의 단백질 데이터를 식별하여 지역 데이터베이스를 동기화하고, 동시에 PPI 네트워크의 데이터도 동기화하여 준다. 본 논문에서는 Uniprot 래퍼(Wrapper)를 사용하여 실시간 동기화를 하였고, 전역 데이터베이스로 UniProt 데이터베이스를 이용하였다.

UniProt 래퍼(Wrapper)는 단백질 아이디, 단백질 이름, 유전자 이름을 입력으로 받아 Uniprot 사이트에서 단백질

요약 정보를 추출하여 제공하는데, 본 논문에서는 단백질 이름, 단백질 엑세션 넘버(protein accession number), 최근 수정 일자, 동의어 정보, 유전자 정보, taxID, GO 정보, 키워드, 서열(sequence)등의 단백질 요약 정보를 추출하도록 개발되었다.

PPI 네트워크에서 특정 단백질을 식별하기 위하여 동기화를 실행하면 웹 프로토콜을 이용하여 UniProt SRS 시스템에 접속되어 선택된 단백질의 정보를 가지고 질의를 작성한다. 작성된 질의가 실행되면 단백질 엑세션 넘버 리스트가 결과로 반환되고, 이 결과값들을 이용하여 UniProt 래퍼(Wrapper)가 반복 실행되어 단백질 요약 정보를 UniProt 데이터베이스로부터 추출한다. 이렇게 추출된 단백질 요약 정보는 XML 형태로 변환되고, 이 XML 형태의 데이터는 래퍼(Wrapper) API에서 DOM 트리를 이용하여 스트링 배열 리스트 형태로 변환되어 PPI 네트워크에 단백질 요약 정보를 넘겨주게 된다. 그림 3 은 위에서 설명한 전역 동기화 과정을 그림으로 보여주고 있다.

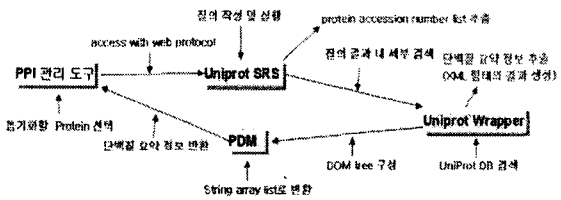

그림 3. 전역 동기화

4. 구현

본 논문에서는 시스템 구현을 위해 HP WorkStation을 사용하고 운영체제로는 Windows 2003 Server를 사용하였다. 사용한 데이터베이스는 Oracle 9i 이며, JBuilder 2005 Enterprise 버전을 사용하여 프로그램 되었다.

선택된 단백질 식별을 위한 지역 동기화는 동기화 컴포넌트에서 수행되는데, 그림 4 는 지역 동기화가 수행된 결과를 나타낸다.

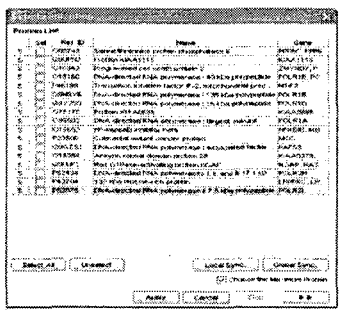

그림 4. 지역 동기화 수행 화면

선택된 단백질 식별을 위한 전역 동기화도 지역 동기화와 마찬가지로 동기화 컴포넌트에서 수행되는데, 수행 과정은 지역 동기화와 동일하다. 그림 5 는 전역 동기화가 수행된 결과를 나타낸다.

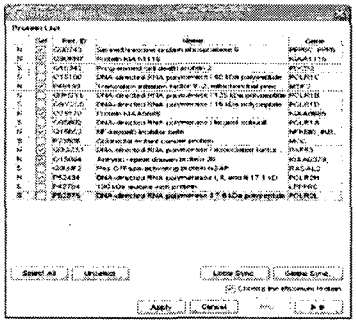

그림 5. 전역 동기화 수행 화면

5. 결론 및 향후 연구 과제

본 논문에서 제안한 래퍼(Wrapper) 기반 단백질 식별은 지역 동기화뿐 아니라 전역 동기화도 가능하도록 하였다. 또한, 실시간 동기화를 통하여 PPI 네트워크에 존재하는 데이터들을 항상 최신 데이터로 유지함으로써 데이터에 대한 신뢰도를 보장하였다.

앞으로는 전역 동기화에 필요한 실시간 동기화 시간을 최소화하기 위한 연구가 진행되어야 한다. 또한, PPI 네트워크를 동적으로 구성하여 지역 동기화의 필요성을 배제할 수 있는 연구도 병행되어야 한다.

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] Swiss-Prot, http://au.expasy.org/sprot/

[4] Gene Bank, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

[5] Z. Lacroix, O. Boucelma, and M. Essid, Tools for integrating and querying web information: The biological integration system, Proceedings of the fifth ACM international workshop on Web information and data management, November 2003.

[6] J. F. Aldana et al., Bioinformatics(BIO) : Integrating Biological Data Sources and Data Analysis Tools through Mediators, Proceedings of the 2004 ACM symposium on Applied computing, 2004.

[7] C. A. Goble, R. Stevens, G. Ng, S. Bechhofer, N. W. Paton, P. G. Baker, M. Peim, and A. Brass, Transparent access to multiple bioinformatics information sources, IBM Systems Journal 40 (2001), no. 2, 532-551.

[8] M. T. Roth and P. Schwarz, A Wrapper Architecture for Legacy Data Sources, Technical report rj10077, IBM, 1997.

[9] UniProt, http://au.expasy.org/