온톨로지 기반의 단백질 반응 데이터 품질향상 연구

장희선*, 원민영*, 김태경**, 조완섭*
* 충북대학교 경영정보학과
** 충북대학교 정보산업공학과
충북 청주시 흥덕구 개신동 충북대학교 학연산 공동기술연구원 906호
sinbi525@gmail.com, dnjsals01@nate.com,{ tkkim, wscho@cbnu.ac.kr}

The Research of PPI Data Quality Improvement on Ontology

Heeseon Jang*, Minyoung Won*, Taekyeong Kim**, Wanseop Cho*
* Chungbuk Univ. Dept. of MIS
** Chungbuk Univ. Dept. of Information Industrial Engineering

요 약

단백질 상호작용(Protein-Protein Interaction : PPI)은 생명체 내에서 생명현상을 유지하기 위한 단백질 간의 유기적인 반응이다. 생물학 실험 도구의 발달로 현재 대량의 PPI 데이터가 발생하고 있으며 이것을 활용하여 생명체를 시스템 관점에서 이해하기 위한 시도가 진행되고 있다. 하지만 현재 PPI 데이터는 중복의 문제, 생물학자들 간의 용어 사용의 불일치성 문제로 인해 데이터의 질이 저하되어 분석결과에 부정적인 영향을 끼치고 있다. 본 논문에서는 현재 인간 PPI 데이터의 문제점을 분석하고 이것을 해결하기 위한 온톨로지 기반의 데이터 정제 방법론을 제시한다. 본 연구를 통해 HPRD내에 존재하고 있던 약 52%의 결함이 있는 데이터를 발견하고 인간 단백질 상호작용 데이터의 품질을 향상시켰다.

1. 서 론

1940년대 이후 분자생물학의 발달로 생명체 내의 핵산과 단백질의 구조를 밝히고 그 분자구조의 특성을 바탕으로 생명현상을 설명하려는 시도가 진행되었다. 그 결과 오랜 시간 실험과 검증과정을 거쳐 대사경로, 조절경로, 신호경로와 같은 생물학적 반응경로 정보가 지속적으로 축적되었다. 최근엔 복잡한 생명체를 시스템 관점에서 생명체를 바라보는 시스템 생물학이 이슈가 되고 있다. 생명체의 경우 부분으로는 생명체 내의 핵산, 단백질 구조 및 반응 메커니즘을 모두 밝히는 것에 대한 한계가 있기 때문에 생명체를 시스템 관점에서 전체적으로 바라보고 이해하는 것이 무엇보다 중요하다는 것이다.

생명체를 전체적인 관점으로 이해하기 위해서는 분자생물학 분야에서 오랜 시간 동안 축적된 대량의 시험 데이터와 현재 지속적으로 발생하고 있는 데이터를 통합하고 정제하는 과정이 필요하다. 과거 실험 데이터는 다양한 포맷으로 데이터 표현에 대한 일정 표준 없이 보관되었다. 그 결과 각 부분에 산재된 데이터를 수집하는 과정이 요구되었으며, 뿐만 아니라 현재 대량으로 쏟아져 나오는 실험 자료들도 일정한 표현 형식으로 하나의 데이터베이스에 수집되는 작업이 필요하다. 이러한 데이터를 모아놓은 대표적인 데이터베이스로 NCBI, DDBJ, KEGG가 있다. 또한, 흩어져 있던 생물학 실험 데이터들이 일정한 양식으로 하나의 데이터베이스에 저장된 후에 데이터가 가지는 의미상의 통합과 정제 과정이 필요하다.

데이터가 대량으로 있어도 데이터의 정확성 및 중복이 많으면 분석 결과에도 부정적인 영향을 끼친다. 특히 현재 생물학자들이 사용하는 생물학 용어의 불일치로 발생하는 문제가 심각하다. 하나의 단백질에 대한 동의어가 수십 개가 검색되기도 하며, 그 결과 동일한 반응에 대하여 서로 다른 표현을 사용함으로써 다른 반응으로 인식된다. 이런 용어의 일관성 문제로 발생하는 문제를 해결하기 위하여 본 논문에서는 온톨로지를 기반으로 인간 단백질 반응 데이터베이스의 데이터 품질개선 방안 및 프로세스를 제시한다. 또한 제시된 문제점들을 기준으로 데이터 정제과정을 거친 실험 결과를 제시한다.

2. 관련연구

본 장에서는 생물학 데이터 정제기법에 대하여 조사하고, 본 논문과 관련 있는 단백질 상호작용 데이터베이스에 대하여 알아본다.

2.1 온톨로지 개요

온톨로지는 단어와 관계들로 구성된 일종의 개념 사전으로 도메인 내에서 공유되는 데이터들을 형식적이고 명백하게 개념화한 규칙이라고 할 수 있다. 이는 도메인에 관련된 단어들이 계층적으로 표현되고, 이를 확장할 수 있는 규칙들을 포함한다. 온톨로지를 활용하면 여러 가지 형식으로 데이터를 저장함으로써 오류를 방지할 수 있다. 온톨로지는 데이터의 공유, 재사용 등을 가능하게 한다.

2.2 생물학 데이터 정제기법

데이터베이스는 데이터베이스 간 통합이 가능하다는 장점이 있다. 그러나 데이터베이스의 범위가 확대되고 복잡도가 증가함에 따라 데이터의 정확성, 중복성, 일관성 등과 같은 데이터 품질 문제가 발생하게 되었고, 이 데이터의 품질관리가 데이터베이스를 운영, 관리하는데 중요한 요구사항이 되고 있다. 단백질 상호작용 정보를 데이터베이스에 저장하고 서비스하는 과정에서 먼저 데이터의 정제 작업이 필요하다. 특히, 이종적인 원본 데이터가 잘 관리되기 위해서는 정제 및 이종성을 제거하기 위한 데이터 변형이 필요하기 때문에 일련의 ETL 정제과정을 거친다. 또한 데이터 중복처리를 위한 방법으로는 일반적으로 정렬과 조인을 이용한다. 이는 간단한 데이터 중복을 제거할 수는 있으나 복잡한 데이터의 중복처리는 해결하기 어려운 단점이 있다.

윈도우 개념을 사용하여 일정한 범위내의 데이터에서 중복을 발견하여 발견 즉시 제거하는 방법을 사용하는 Bitton's method[1]의 방법이 있다. 그리고 이와 같은 데이터 정렬을 이용한 기법 외에 Intelliclean[1]과 같은 전문적인 지식을 기반으로 하여 데이터 정제 툴을 사용하는 방법이 있다. 이는

자동화 툴만으로 해결할 수 없는 오류에 대해 전문적인 지식을 기반으로 개선하는 방법이다.

이 논문에서 제시된 데이터 정제기법은 지식 기반데이터 정제 방법으로서, 먼저 불규칙한 데이터를 규칙에 맞게 정제하고, 그 후 전문가 및 사용자의 지식과 이들의 검증을 통해 데이터를 정제하는 방법이 사용되었다. 본 논문은 논리적인 레벨과 물리적인 레벨을 분리하여 데이터를 정제하고 논리적인 레벨에서는 매핑, 뷰 매칭, 클러스터링, 그리고 병합 논리연산자를 사용하여 데이터 정제 과정을 거친다.

2.3 단백질 상호작용 데이터베이스

실험도구의 발달로 단백질 상호작용 데이터가 대량으로 쏟아져 나오고 있다. 이 데이터들은 다음과 같이 여러 데이터베이스에 저장되어 서비스되고 있다.

① BIND (Biomolecular Interaction Network Database)[2]: 단백질 상호작용 정보와 상호작용의 경로정보를 함께 제공한다. interaction, molecular complex, pathway records 등이 포함되어 있다.

② DIP (Database of Interacting Proteins)[3]: 두 단백질의 상호작용에 대한 데이터베이스로 웹에서 공개적으로 사용이 가능하고, 단백질 상호 작용, 신호경로, 다중 반응, 복잡계 시스템을 제공하는 것을 목적으로 한다.

③ MINT (Molecular Interactions Database)[4]: biological molecules(proteins, RNA, DNA)의 기능상 상호작용 데이터를 저장하는 데이터베이스이다.

④ HPID (The Human Protein Interaction Database)[5]: 기존의 정보와 실험을 통해 미리 계산된 인간 단백질 상호작용 정보를 제공하고, 사용자에게서 받은 단백질들 사이의 가능한 상호작용을 예측하고 마지막으로 새로운 인간 단백질 상호작용 데이터를 사용자들이 직접 입력한다.

3. 단백질 상호작용 데이터 정제기법

본 장에서는 실험에 사용된 인간 단백질 반응 데이터에 대하여 알아보고, 이 데이터의 질을 향상시키기 위한 기법을 제안한다.

3.1 인간 단백질 반응 데이터

HPRD(Human Protein Reference Database)는 인간 단백질 내의 단백질 상호작용 정보를 여러 관점에서 통합적으로 관리하는 데이터베이스로써, 모든 데이터는 생물학 정보를 제공하는 다른 데이터베이스로부터 참조하여 서비스하고 있다. 또한 새로운 단백질 정보는 관련 생물학 논문을 읽고 발췌하는 작업으로 생물학자들이 직접 확인 작업을 거친 후 입력되어 구축되고 있다. 하지만 이기종 데이터베이스로부터 데이터를 가져오는 작업과 수작업에 의한 데이터 저장방법에 따라 여러 문제점이 발생하여 데이터 품질을 저하시키고 있다.

3.2 인간 단백질 반응 데이터의 문제점

데이터 값의 정확성에 대한 품질을 평가할 때는 값의 완전성, 일관성, 정확성에 관점을 두고 평가를 한다. 완전성이란 요구되는 데이터의 값을 보유하고 있는가를 의미한다. 데이터 내용의 정확성은 수록되어 있는 데이터가 오류 없이 원본 데이터가 가지고 있는 값과 동일한가를 의미하며, 일관성은 데이터베이스의 관련 있는 데이터 값들이 상호모순 없이 일관되어야 함을 의미한다.[6]

데이터 품질 평가 모델로 HPRD를 평가해보았을 때, HPRD는 수작업을 통한 단백질 데이터 수집방법으로 인해 논문 내에 존재하는 단백질 데이터의 누락과 비체계적이고 표준화 되지 않은 학명을 사용하고 있다. 이와 더불어 HPRD 내에 존재하고 있는 protein정보를 누락시키고 있는 문제점(true-negative)을 포함하고 있다. 그리고 이기종 데이터베이스로부터 가져오는 데이터의 포맷이 달라 데이터 통합에 문제점이 발생한다. 따라서 이러한 문제점 때문에 실제로는 오랜 시간을 투자하여 얻은 단백질 상호작용 정보를 제대로 서비스하고 있지 못하다는 사실을 발견했다.

HPRD 인간 단백질 반응 네트워크는 크게 다음 5가지의 문제점을 가진다.

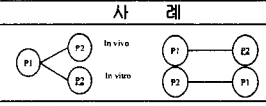
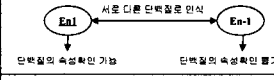
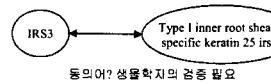
첫째, 데이터의 중복을 가지고 있다. 이는 중복 데이터의 유일성을 분석하여 재설정하는 작업이 필요하다.

둘째, 동의어 처리가 분명하지 않다. 생물학 데이터는 다양한 형태로 존재하는데, 다양한 생물학 분야에서 생성되는 지식 도메인은 서로 중복되어 존재하거나 서로 보완적인 타입으로 존재하기도 하며, 전문용어를 갖기도 한다.

셋째, 원본데이터 포맷오류를 가지고 있다. 생물학 데이터의 명명법은 이명법에 기초하고 있지만, 일관성 있는 규칙의 정확한 포맷은 없다. 하지만 데이터베이스에 입력과 검색을 위해서는 일관성을 보장하는 데이터 포맷이 요구되고 유지되어야 한다. 그리고 명명규칙의 일관성 미흡, Protein명 부여규칙 정의 필요성 및 규칙을 준수해야 하는 문제점을 가지고 있다.

넷째, 데이터 검색 시 세부 데이터가 불일치한다. 가장 많은 해결 시간을 필요로 하는 문제로서, 일부분에 대해서만 일치하는 단백질 데이터가 질의를 먼저 얻고자하는 결과와 일치하는 대한 정확성을 판별할 수 있는 생물학자들의 검증이 요구된다.

마지막으로 핸드코딩에 의한 데이터 입력방법에 의해 오류가 발생한다. 직접 논문을 읽고 데이터를 수집하는 방법은 데이터에 대한 100% 리뷰가 곤란할 뿐 아니라, 많은 리소스를 요하는 프로세스다. 하지만 가장 큰 문제는 단백질 데이터는 Tree Form Text이기 때문에 표준화된 접근이 불가능 하며, 많은 시간을 필요로 하기 때문에 데이터의 업데이트 또한 부정확하고 빠르지 못하다. 한 가지 예로, 'Fodrin' 단백질은 현재 'Fodrin alpha' 와 'Fodrin beta' 두 종류가 있지만, 논문을 읽고 데이터를 수집하는 방식으로 인해 논문을 미처 읽지 못하고 발견하지 못하였다면 데이터베이스 내에 존재하지 않는 데이터로 결과가 도출된다. 반자동화 기법을 통해 온톨로지 기반 단백질 반응 데이터를 수집하는 작업을 거치면 보다 정확하고 빠르게 현존하는 데이터를 입력할 수 있을 것이라 기대된다.

종 류	사 례
데이터의 중복	
동의어 처리 문제	USP8, UBPY, HUMORF8, Ubiquitin thiolesterase8, Deubiquitinating enzyme humor f8 이 모두 같은 단백질을 의미함
원본 데이터 포맷 오류	
데이터 검색시 세부 데이터 불일치	
핸드코딩에 의한 오류 발생	- 신속하지 못한 데이터의 입력 - 잘못된 단백질 코드형식으로 입력

[표1]HPRD 인간 단백질 반응 네트워크 문제점

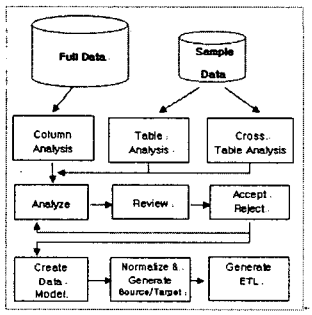
3.3 인간 단백질 반응 데이터 정제기법

본 논문은 HPRD에서 서비스되고 있는 인간 단백질 데이터를 온톨로지와 자동화 및 반자동화 데이터 정제기법을 사용하여

단백질 반응 데이터 품질을 향상시켜 보다 의미 있고 정확한 정보를 제공하고자 한다. 본 논문에서는 지식 기반의 데이터 정제과정을 사용하여 단백질 데이터 정제를 한다. 일반적으로 대부분의 데이터 정제과정은 데이터 변환, 제약조건 실행, 중복제거에 집중되어 있다. 이러한 방법에 기초하여 HPRD 데이터를 정제하는 프로세스는 다음과 같다.

① Data Profiling(데이터 프로파일링)

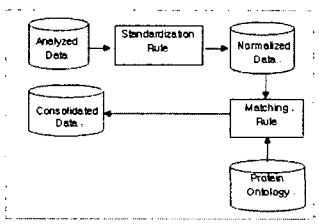
데이터 프로파일링이란 데이터 자체로부터 지식을 획득하는 프로세스로서 데이터 품질 관리의 전단계라 할 수 있다. 모든 통합과정은 데이터 프로파일링 과정에서 시작되며, 데이터 값의 누락, 도메인의 일관성, 관련 항목간의 상관성 등을 보기 쉽게 나타내준다[6]. 단백질 반응 데이터 품질을 향상시키기 위한 본 연구는 [그림1]과 같이 이루어졌다. 먼저 표본데이터 집합을 만들어 교차분석과 테이블분석을 통해 단백질 데이터의 전반적 이해와 더불어 유효하지 않은 값과 결함 데이터를 발견하고, 기본 값을 파악하는 등 데이터 규칙을 찾아내어 전체 데이터에 대해 분석 및 검사 과정을 거친다. 이러한 소스 분석과정을 거친 후 리뷰작업을 통해 올바르지 못한 데이터는 재분석 작업으로 피드백 되며, legacy 데이터에서 새로운 데이터 모델을 만들어 ETL 작업 정의를 생성한다.



[그림 1] 데이터 품질을 향상을 위한 실험 과정[7]

② Data Cleansing(데이터 정제)

프로파일링 과정을 통해 분석된 데이터로부터 존재하고 있던 형식의 불일치와 중복되어 있는 데이터를 찾아 이를 단일화하기 위한 데이터 정제과정을 필요로 한다. 일관적인 데이터로 변환하기 위한 표준규칙을 발견하여 프로파일링 단계를 거친 데이터를 표준규칙에 적용하여 정규화된 데이터로 변환한다. 그리고 정규화된 데이터이기는 하나 또는 그 이상 중복된 데이터는 온톨로지에서 찾은 다음 매칭 규칙에 따라 유일성이 보장되는 단일화된 데이터로 변환한다. 이와 같은 단계를 반복적으로 수행함으로써 데이터가 가지고 있는 결점을 줄여간다.



[그림 2] 온톨로지 기반 데이터 정제 과정

4. 실험결과

2장에서 설명한 지식 기반의 데이터 정제기법을 기반으로 HPRD의 단백질 상호작용 데이터의 문제점을 줄이고자 했다. HPRD의 데이터 55,604건을 오라클 데이터베이스로 옮겨와 먼저 HPRD내에 존재하는 데이터를 검증한 결과 데이터베이스에 존재하고 있지만, 존재하지 않는 것으로 되는 것으로 잘못 판명되어 제공되는 데이터는 6,194건으로써, 다음 표와 같다.

1차 정제 과정	HPRD 단백질 데이터 총 개수	55,604
	분석 대상 데이터 수	6,194 (모든 문제유형 포함)
1차 정제 결과		6,194의 분석데이터 중에서 44%(2,733)의 데이터 재정의

[표2] HPRD 1차 데이터 정제 결과

1차 정제 과정을 거친 단백질 데이터는 다시금 중복을 제거하는 과정을 필요로 한다. P1-P2/P1-P2의 동일한 단백질 상호작용 형태로 중복된 데이터를 제거하는 단계를 수행한다. 그리고 단백질 반응은 방향성을 가지고 있지 않기 때문에 P1-P2와 P2-P1은 동일한 단백질 반응이 된다. 이에 따라 다시 P1-P2/P2-P1의 형태로 중복된 단백질 반응 데이터를 제거하여 유일성을 갖는 단백질 데이터로 재구성을 한다.

2차 정제 과정	P1-P2 / P1-P2	8,290
	P1-P2 / P2-P1	20,633
2차 정제 결과		55,604 개의 데이터 중 총 52%의 데이터 중복 발견

[표3] HPRD 2차 데이터 정제 결과

5. 결론 및 향후 연구

현재는 생물학 데이터에 대한 일관성 있는 단일화된 규칙이 없기 때문에 완벽한 데이터 정제과정 및 평가는 불가능하다. 따라서 본 논문은 생물학자의 도움을 받아 일정한 규칙을 갖는 데이터 정제 과정을 정의하고 보다 향상된 인간 단백질 반응 데이터로 개선하고자 했다. 하지만 완벽한 단백질 데이터 품질을 위한 정제 방법 및 평가방법에는 한계가 있으며, HPRD에 국한되어 있다는 한계점이 있다. 향후 연구는 모든 단백질 결함 데이터를 찾기 위한 Free Form Text 데이터에 대한 정제 기술을 개발해야 하는 것이다. 또한 수작업을 통한 데이터 수집 과정을 text mining을 이용한 자동화 수집 과정을 추진하는 것이다. 현재 추진 중인 과정으로서 수작업을 간소화하여 빠르게 증가하는 단백질 데이터를 빠르게 업데이트 할 수 있을 뿐만 아니라, 데이터 수집 이전에 데이터에 대한 이해가 가능하게 되며, 결함 데이터의 사용으로 인한 불확실성을 보다 많이 제거 가능하도록 해야 한다.

참고문헌

[1] Katherine Grace Herbert et al., "New techniques for improving biological data quality through information integration", Department of Computer Science, May 2004

[2] Biomolecular Interaction Network Database, http://www.bind.ca/Action

[3] Database of Interacting Proteins http://dip.doe-mbi.ucla.edu/

[4] Molecular Interactions Database http://160.80.34.4/mint/

[5] Mammalian protein-protein interaction database(PPI), http://wilab.inha.ac.kr/hpid

[6] 김문영:"체계적인 데이터 품질 관리 대안 찾아라",ZDnet Korea, http://www.zdnet.co.kr

[7] 이용우:"데이터 품질관리 자동화 방안", 2004데이터베이스그랜드컨퍼런스,2004