RDF(S) 모델에 기반한 다양한 형태의 단백질 데이타베이스 통합

(Integration of Heterogeneous Protein Databases Based on RDF(S) Models)

이 강 표 † 유 상 원 † 김 형 주 ††
(Kangpyo Lee) (Sangwon Yoo) (Hyoung-Joo Kim)

요 약 현재 생물학 분야에는 단백질이라는 동일한 대상에 대해 각기 고유한 의미를 지니고 있는 다양한 형태의 단백질 분석 데이타베이스들이 존재한다. 이렇게 산재되어 있는 이종의 단백질 정보들을 효과적으로 통합한다면 개개의 데이타베이스로부터는 얻을 수 없는 유용한 정보를 도출해낼 수 있다. 생물학 데이타의 특성상 이 각각의 정보들은 자신만의 고유한 형태와 의미를 지니는데, 시맨틱 웹 기술의 표준인 RDF(S) 모델을 이용하여 데이타를 기술하면 형태론적인 통합뿐만 아니라 의미론적인 통합까지 이루어낼 수 있다. 이에 본 논문에서는 RDF 통합 스키마에 기반한 새로운 통합 레이어(layer)를 제안하였다. 이를 위해 개념적 모델 차원으로서는 단백질 정보를 중심으로 통합 스키마를 구축하였고, 표현적 모델 차원으로서는 래퍼(wrapper)가 해당 데이타베이스들로부터 필요한 정보를 취하여 동적으로 RDF 인스턴스를 구축하는 방법을 제안하였다. 실제로 이 통합 레이어는 연구자들이 필요로 하는 통합 질의 예제를 성공적으로 처리하여 그 결과를 보여줄 수 있음을 확인하였다.

키워드 : RDF(S), 데이타 통합, 단백질, 생물정보학

Abstract In biological domain, there exist a variety of protein analysis databases which have their own meaning toward the same target of protein. If we integrate these scattered heterogeneous data efficiently, we can obtain useful information which otherwise cannot be found from each original source. Reflecting the characteristics of biological data, each data source has its own syntax and semantics. If we describe these data through RDF(S) models, one of the Semantic Web standards, we can achieve not only syntactic but also semantic integration. In this paper, we propose a new concept of integration layer based on the RDF unified schema. As a conceptual model, we construct a unified schema focusing on the protein information; as a representational model, we propose a technique for the wrappers to aggregate necessary information from the relevant sources and dynamically generate RDF instances. Two example queries show that our integration layer succeeds in processing the integrated requests from users and displaying the appropriate results.

Key words : RDF(S), Data Integration, Protein, Bioinformatics

· 본 연구는 정보통신부 및 정보통신연구진흥원의 대학 IT연구센터 육성지원사업(IITA-2006-C1090-0603-0031)의 연구결과로 수행되었음

† 학생회원 : 서울대학교 컴퓨터공학부
kplee@idb.snu.ac.kr
swyoo@idb.snu.ac.kr
†† 종신회원 : 서울대학교 컴퓨터공학부 교수
hjk@snu.ac.kr

논문접수 : 2006년 11월 2일
심사완료 : 2007년 12월 8일

정보과학회논문지: 데이타베이스 제35권 제2호(2008.4)

1. 서 론

최근 생물정보학의 비약적인 발전으로 인해 생물학 각 분야에서 연구된 수많은 데이타들이 각기 고유한 데이타베이스에 저장되어 있다. 특히 생명체의 주요성분인 단백질이 세포의 주요 구성물질이며 생명현상과 밀접한 관련이 있다는 근거 하에 단백질체학(proteomics)에 대한 연구가 활발히 진행되고 있으며, 이와 관련된 많은 데이타베이스들이 존재한다. 이러한 단백질 분석 데이타베이스(proteome analysis database)는 하나의 단백질을 바라보는 관점에 따라 다양한 정보를 지니게 되는데, 그림 1과 같이 이렇게 다양한 관점의 단백질 정보들을

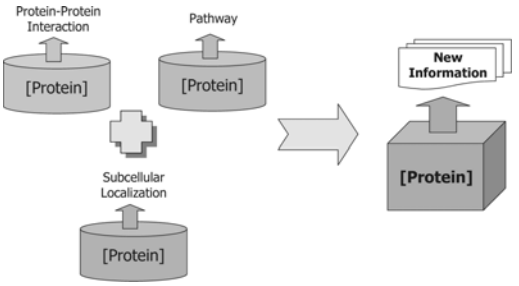


그림 1 다양한 관점의 단백질 분석 데이타베이스의 통합

효과적으로 통합한다면 개개의 독립적인 정보로부터는 얻을 수 없는 새롭고 유용한 정보를 도출해 낼 수 있다. 가령, 어떤 데이타베이스에서의 예측값이 실제로 맞는지를 다른 데이타베이스를 통해 확인해 볼 수도 있으며, 혹은 어느 한 데이타베이스에서는 알 수 없는 부족한 정보를 다른 데이타베이스를 통해 알아낼 수도 있는 것이다. 이와 같은 맥락에서 최근 신진 대사 회로와 단백질 상호 작용 네트워크의 중요성에 대한 연구가 광범위하게 진행되고 있는데[1,2], 신진 대사 회로나 세포 신호와 같은 서로 다른 생물 조직간 상호 정보를 이해할 수 있다면 실제 생물학적인 현상을 모델링하고 이해하는데 큰 도움이 될 수 있다는 연구 결과가 있다[3].

생물학 데이타는 생물학 도메인의 특성상 자신만의 고유한 형태(syntax)와 의미(semantics)를 지닌다. 특히 의미론적인 측면에서 바라볼 때 각각의 생물학 데이타는 복잡하고도 미묘한 의미를 지니며, 생물학 데이타를 통합하고자 할 때 이러한 의미들을 고려하지 않을 수 없다. 그러나 형태론적인 통합에 치중하였던 기존의 전통적인 통합 기법으로는 이러한 의미들을 제대로 반영하는 것이 쉽지 않다. 특히 단백질 정보의 경우에 있어서 전통적인 통합 기법은 그 한계를 보일 수밖에 없는데, 이는 단백질 정보를 다양한 관점에서 바라보는 여러 데이타베이스들을 단순히 형태론적으로 통합하였다고 해서 우리가 원하는 다양한 복합 질의에 대한 해답을 얻을 수는 없기 때문이다. 따라서 생물학 데이타들 간의 특별한 의미를 고려할 수 있는 새로운 통합 기법의 필요성이 대두되었다.

RDF(Resource Description Framework)[4]는 월드 와이드 웹(world wide web)의 자원들에 대한 정보를 표현하기 위한 언어로서, W3C(World Wide Web Consortium)[5]에 의해 시맨틱 웹(semantic web)을 기술하기 위한 표준으로 제정되었다. RDF는 생물학 도메인을 표현하고 통합하는 데 있어서 적합한 언어라고 할 수 있는데, 이는 다음과 같은 RDF의 특성들에 기인한다.

(1) RDF는 표현력이 우수하다. RDF는 웹 상에 존재하는 자원들과 그들 간의 관계를 정의하기 위해 탄생한 언어이기 때문에, 단백질 정보의 복잡한 형태와 의미를 충분히 반영할 수 있다. 이에 반해 기존의 관계형 데이타베이스나 XML은 의미를 제대로 반영하지 못하고, 지식을 표현하고 추론하기 위한 표현력이 부족하기 때문에 생물학 도메인을 표현하는 데 있어서 그 한계를 드러낸다고 할 수 있다. 이와 같은 우수한 표현력으로 인해 현재 Gene Ontology (GO)[6], NCI thesaurus[7], Uniprot[8] 등과 같은 생물학 도메인에서 RDF로 데이타를 기술하고 있다.

(2) RDF는 데이타 상호운용성(interoperability)이 뛰어나다. 현재 공개되어 있는 수많은 생물학 데이타들은 웹 상에 산재되어 있으며 대부분 이종의(heterogeneous) 정보들이다. RDF는 트리플(triple) 형태라는 공통의 데이타 모델을 제공함으로써 이종의 데이타 통합을 위한 훌륭한 환경을 제공해준다. subject, predicate, object로 이루어진 RDF 트리플을 통해 단순하고 유연한 공통의 프레임워크(framework)를 가능하게 해주는 것이다. 또한, 자원들의 클래스와 속성에 대한 공통의 RDF 어휘를 정의하기 위해 고안된 RDF 스키마(RDFS)[9]를 이용하면 생물학 데이타에 대한 스키마 정의도 기술할 수 있다.

(3) RDF는 열린 세계 가정(open world assumption)에 기초하고 있기 때문에 유연한 모델이다. 열린 세계에서는 명시적으로 진술된 사실만이 참(true)이다. 명시적으로 진술되지 않았다고 해서 거짓(false)이라고 할 수는 없으며, 단지 알려지지 않은 것(unknown)이 된다. 즉, 어떠한 새로운 진술도 기존의 참인 진술을 거짓으로 바꿀 수 없다는 의미로서, 새로이 추가되는 진술이 기존의 데이타 모델을 깨트릴 수 없게 된다. 이는 다수의 데이타베이스 스키마를 추가해 가면서 통합하는 데 있어서 큰 장점이 될 수 있다. 반면, 기존의 SQL이나 XML은 닫힌 세계 가정(closed world assumption)에 기초하고 있으며 새로운 진술이 기존의 참인 데이타 모델을 깨트릴 수 있기 때문에 새로이 데이타 스키마가 추가될 때 여러 가지 사항들을 고려해야 하는 불편함이 따른다.

본 논문의 구성은 다음과 같다. 2장에는 RDF에 기반한 생물학 데이타 통합과 관련된 기존의 연구를 소개한다. 3장에서는 본논문에서 제안하는 통합 레이어에 대하여 개념적 모델과 표현적 모델이라는 두 차원에서 상세하게 논의한다. 4장에서는 통합 레이어의 생물학적인 실제 활용과 기존연구와의 차별성을 논의한다. 끝으로 5장에서는 본논문에서 제안하는 통합 레이어의 장점과 공헌, 그리고 향후 행해질 연구 방향에 대해서 논의한다.

2. 관련 연구

생물학 데이타베이스를 통합하려는 시도는 현재 활발하게 이루어지고 있다. 서론에서 밝힌 바와 같이 수많은 생물학 데이타베이스들이 산재해 있는 현 상황에서 데이타의 통합은 필수적이라고 할 수 있기 때문이다. [10]은 현재 진행 중에 있는 다양한 시도들을 잘 정리하고 있다. 여기에 소개된 통합 시스템들을 여러 관점에서 정리해보면 표 1과 같다. 이 통합 시스템들은 각기 다른 형태로 산재되어 있는 생물학 데이타베이스들을 통합할 때 발생하는 문제점들을 공통적으로 다루고 있다. 이 문제점들은 결국 데이타베이스들 간의 이질성에 기인하는데, 이 이질성을 극복하기 위해 표 1과 같이 서로 다른 접근방법을 취하고 있다.

또한 [11]에서도 생물학 분야에서 발생할 수 있는 다양한 통합의 이슈들을 다루고 있다. 특히 데이타 통합을 위한 설계방법을 데이타 웨어하우스(data warehouse), 데이타 연합(data federation), 중개스키마를 이용한 데이타 연합(data federation with mediated schema), 동료 데이타관리시스템(peer data management system)의 네 가지로 상세히 구분하여 각각의 장단점을 자세히 소개하였다. 뿐만 아니라, 위의 네 가지 구분에 근거하여 [10]에서 소개되고 있는 시스템들을 포함한 다양한 통합 시스템들을 비교, 분석하였다.

한편, 최근 RDF를 이용하여 다양한 형태로 존재하는 생물학 데이타를 통합하려는 시도가 시작되고 있다. 그 중 YeastHub[12]는 이종의 지놈(genome) 데이타들을 통합하는 데 있어서 시맨틱 웹 기술을 어떻게 활용할 수 있는지를 보여주기 위해 효모(yeast) 정보를 프로토타입으로 삼아 접근하였다. YeastHub는 효모 정보를 제공하는 여러 원본 데이타베이스들로부터 얻어진 데이타들을 정해진 포맷의 RDF로 변환하여 데이타 웨어하우징(data warehousing) 형태로 관리한다는 점이 특징이다. 원본 데이타베이스들의 형태는 현재 가장 많이 이용되고 있는 일반 파일(tab-delimited flat file)과 RDF로 제한하였다. 이 두 가지 형태의 원본 데이타를 대상으로, 메타 데이타는 RSS(RDF Site Summary)[13]라 불리는 웹 컨텐트에 대한 규격을 빌려와 정형화된 RDF 형식으로 변환하였고, 일반 데이타는 원본의 스키마에 충실하게 RDF로 변환하였다. 변환된 RDF는 RDF 데이타 저장/질의 시스템인 Sesame[14]에 데이타 웨어하우스로 구축되어, 사용자로 하여금 RQL과 SeRQL[15] 질의를 허용하도록 하고 있다. 그러나 웨어하우징 기법을 채택한 YeastHub는 빠르게 변화하고 추가되는 생물학 데이타의 특성을 고려해볼 때 원본 데이타와의 동기화 문제를 해결해야 하는 문제를 안고 있으며, 키 값을 통한 단순한 조인만을 제공한다는 단점이 있다.

한편, BioDash[16]는 신약 개발을 위해 필요한 관련 정보들을 하나로 묶어 직관적이고 상호작용 가능한 시각

표 1 대표적인 생물학 데이타베이스 통합 시스템

	SRS	K2/ Bio-Kleisli	TAMBIS	Discovery-Link	BACIIS	Bio-Navigator	GUS	KIND	Entrez
통합의 목표	Portal, browsing-based	Query-oriented	Query-oriented	Query-oriented middleware	Query-oriented	Portal, browsing-based	Query-oriented	Query-oriented	Portal, browsing-based
데이타 모델	Linked text records	Semi-structured, object-oriented	Structured, object-relational	Structured, object-relational	Structured, object-relational	Text model	Structured, relational	Semi-structured, object-oriented	Linked text records
대상 데이타베이스 포맷	Files or databanks with structured text	RDB, formatted text files, & binary files	Tools, processes, & proprietary flat file structures	RDB & flat files	Data source schemas	Text, HTML, & XML	RDB	Various types	RDB & other various types
통합 방식	Navigational	Mediator-based	Mediator-based	Mediator-based	Mediator-based	Navigational	Data warehousing	Mediator-based	Navigational
주요 특징	Keyword-based retrieval	Loosely-coupled, CPL (Collection Programming Language)	Source-independent ontology	Query optimization	The extensive use of domain knowledge base (KB)	User-defined preferred execution path	Data filtering & annotations	Using formal ontologies	Web-based, link-driven

화를 제공해주는 시스템으로서, YeastHub와는 대조적으로 필요한 정보를 병합하여 요구 발생시마다 제공해주는(on-demand retrieval) 방식을 채택하고 있다. 이를 위해 BioHaystack이라는 시맨틱 웹 브라우저(semantic web browser)를 제안하고 있는데, 이는 RDF나 OWL 등과 같은 시맨틱 웹 문서들을 읽어 들인 후 이를 취합하고 가공하여 기존의 웹 브라우저와 유사한 모습으로 시각화하여 화면에 제시해주는 새로운 개념의 도구이다. 아울러 BioHaystack이 데이타를 취합, 가공하여 사용자가 원하는 정보만을 제공해주기 위해 시맨틱 렌즈(semantic lens)라는 지능형 컴포넌트도 제안하였다. 시맨틱 렌즈는 필요한 정보들을 걸러낸 후 이를 Adenine이라는 언어로 정의하여 BioHaystack에 전달해주는 역할을 한다. 이처럼 BioDash는 데이타 통합을 이루었을 때 얻어지는 시각화 측면의 장점에 주력하였으며, 특히 화합물 대사 회로와 분자 대사 회로를 병합하여 예상치 못했던 새로운 정보를 눈으로 발견할 수 있었던 실험이 주목할 만하다. 하지만 원본 데이타가 RDF나 OWL로 표현되어 있어야지만 BioHaystack에서 인식이 가능하기 때문에 그 밖의 이종의 형태로 기술된 데이타에 대해서는 처리를 할 수 없다. 또한 유연한 질의가 아닌 단순한 브라우징만을 제공한다는 단점도 있다.

이와 같이 여러 생물학 데이타베이스들을 RDF를 통해 통합하려는 시도는 이제 시작 단계에 불과하다. 이에 우리는 앞서 살펴본 시스템들을 개선하여, 생물학 데이타의 특성을 온전히 반영하면서 이종의 데이타베이스들로부터 최신의 정보를 유연하게 제공할 수 있는 통합 시스템을 제안할 필요를 느끼게 되었다. 우리는 본 논문을 통해 데이타 동기화를 고려할 필요가 없으며, 이종의 데이타 형식을 지원하며, 단순한 조인만이 아닌 Is-a 등과 같은 의미론적 질의도 제공할 수 있는 통합 시스템을 제안하였다.

3. 통합 레이어(INTEGRATION LAYER)

본 논문에서는 사용자부터 통합 질의를 입력 받으면 복수의 원본 데이타베이스들로부터 필요한 정보를 취합하여 이를 RDF로 변환한 후, 사용자에게 최종 결과를 전달해주는 일련의 과정을 처리하는 RDF 기반의 통합 레이어를 제안한다.

3.1 OASIS 시스템과 통합 레이어

본 논문에서 제안한 통합 레이어는 OASIS(Omics AnalySIS)[17]라는 단백질 분석 시스템의 일부 컴포넌트로 구현되었다. 현재 OASIS 시스템은 하나의 단백질을 바라보는 서로 다른 관점을 지닌 4개의 데이타베이스로 구성되어 있다. 첫 번째는 Gene Ontology[6] Annotation Database(이하 GOA DB)이다. Gene Ontology (GO)는 생물학적인 개념들을 나타내는 표준 어휘 체계라고 할 수 있는데, GOA DB는 GO에서 제공하는 주석 정보를 담고 있는 데이타베이스이다. 두 번째는 Protein-Protein Interaction[18] Database(이하 PPI DB)이다. PPI DB는 단백질 간의 상호작용을 분석하여 그 연관성을 알아냄으로써 알려지지 않은 단백질의 기능을 유추하거나 단백질들 간의 새로운 관계를 예측할 수 있는 정보를 제공해준다. 세 번째는 KEGG[19] Pathway Database(이하 Pathway DB)이다. 대사회로(pathway)란 생물체 내에서 에너지의 방출과 획득, 그리고 그와 관련된 부산물들의 연결 관계를 표현한 것인데, 이 정보를 통하여 대사산물의 특성을 밝혀 특정한 화학 물질을 생산하는 데 활용될 수 있다. 이중 KEGG는 공개된 관련 데이타베이스 중 가장 많이 이용되고 있는 대표적인 대사회로 데이타베이스이다. 네 번째는 Subcellular Localization[20] Database(이하 SL DB)이다. SL DB는 특정 유전자가 세포 내 어떤 위치에서 단백질로 생성될 것인지를 예측하는 정보를 담고 있다. 단백질의 위치를 알면 그 기능에 대한 예측에 근거를 제시해줄 수 있다. 이와 같이 OASIS의 4개의 데이타베이스들은 제각각 단백질의 다양한 측면의 정보를 다양한 형태로 저장하고 있음으로써, 데이타 통합을 위한 훌륭한 환경을 제공해주고 있다.

그림 2는 전체 OASIS 시스템의 구조를 간략히 보여준다. 4개의 원본 데이타베이스들은 OASIS 시스템의 최 하단부에서 작동하고 있고, 통합 레이어는 이들의 상단에 위치한다. 그리고 이 통합 레이어에는 RDF 통합 스키마(unified schema)와 각 원본 데이타들과 통신하는 래퍼(wrapper)들이 위치한다. 통합 스키마와 래퍼에 대해서는 이어지는 섹션에서 자세히 다루겠다. 말단 사용자는 각각의 원본 데이타베이스의 스키마에 대해서는 인지할 필요가 없으며, 미리 구축된 RDF 통합 스키마만을 바라보고 통합 레이어에 질의를 요청한다.

3.2 개념적 모델: 통합 스키마 구축

개념적 모델로서 통합 스키마의 구축은 통합 레이어

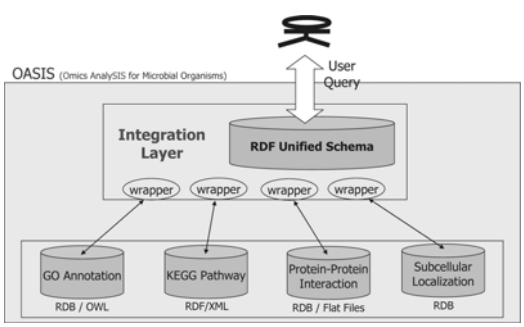


그림 2 OASIS 시스템 구조

의 핵심이라고 할 수 있다. 여기서 우리는 통합 스키마를 '각각의 원본 데이타베이스의 스키마들을 어떤 의미를 기준으로 통합하여 단일화한 가상의 스키마'라 정의한다. OASIS 환경에서 기준이 되는 의미는 바로 단백질이다.

통합 스키마를 구축하기 위해서는 각 원본 데이타베이스의 스키마를 파악하는 작업이 가장 먼저 선행되어야 한다. 그림 3은 원본 데이타베이스의 스키마를 각각 분석하여 이를 다이어그램으로 표현한 것이다. (GOA DB, PPI DB, SL DB는 관계형 모델로 표현되었고, Pathway DB는 그래프 모델로 표현되었다.) 여기서 GOA DB의 gene_product 테이블, PPI DB의 protein 테이블, SL DB의 protein 테이블, 그리고 Pathway DB의 entry 노드에 주목해보자. 이들은 서로 다른 고유한 값을 가지고 있지만 모두 공통적으로 단백질 정보를 의미하고 있다. 따라서 이들 테이블 혹은 노드가 4개의 데이타베이스를 통합하는 기준 축(pivot)이 될 수 있다. 각 데이타베이스들이 단백질 정보라는 의미를 공유함으로써 개념적 모델 상에서 통합되고 있다는 점에서, 의미를 고려한 이 기준 축의 선정은 통합 스키마의 구축에 있어 핵심적인 과정이라고 할 수 있다. 우리는 생명공학 연구원들의 요구사항을 반영하여 OASIS 환경에서 단백질이라는 의미를 기준 축으로 선정하였다.

단백질 정보를 기준으로 통합한 통합 스키마를 RDF 그래프로 표현한 모습은 그림 4와 같다. 그림에서 보는 바와 같이 각각의 원본 스키마에서 단백질 정보를 나타내는 엔터티(entity)들이 통합 스키마에서는 Protein이라는 하나의 클래스로 통합되었다. Protein 클래스의 속성들을 살펴보면 NCBI[21] 정보(ncbi_gi, ncbi_geneid), Pathway DB의 정보(dblinks_id), Uniprot[8] 정보(uniprot_acc), GO[6] Term 리스트 등 원본 스키마에서 단백질 정보를 지니고 있던 엔터티에 접근하기 위한 ID 값들을 고스란히 지니고 있다. 이는 다음 섹션에서 설명할 래퍼가 원본 데이타베이스에 접근해야 할 때 Protein 클래스로부터 필요한 정보를 얻어오기 위함이다. 그림에서 PPI 클래스는 원본 PPI DB의 interaction 테이블이 통합 스키마에 포함된 모습이고, SL 클래스는 원본 SL DB의 subcellular_localization 테이블이, Term과 GeneProduct, GraphPath 클래스 등은 원본 GOA DB의 term, gene_product, graph_path 테이블 등이 통합 스키마에 포함된 모습이다. 한편, Pathways, Entry, Reaction 클래스는 Pathway DB의 Pathway, Entry, Reaction 노드와 그에 딸린 속성들이 통합 스키마에 포함된 모습이다. 그림에서 볼 수 있듯이 원본 스키마의 모든 엔터티들이 통합 스키마에 포함된 것이 아니라, 통합 스키마에서 필요로 하지 않거나 다른 원본 스키마의 엔터티와 중복되는 경우에는 통합 스키마에서 생략되었다. 따라서 통합 스키마는 사용자의 통합 질의를 수행하기 위해 필수적인 엔터티들만 모아 하나의 단일 스키마를 형성한 것이라 볼 수 있다. 이렇게 구축된 통합 스키마는 RDF(S)로 기술되어 OASIS 시스템에 저장된다. 그림 5는 RDF로 기술된 통합 스키마의 일부로서, Organism, Protein, PPI 세 개의 클래스들의 속성과 그들 간의 관계가 어떻게 RDF로 기술되어 있는지를 보여준다.

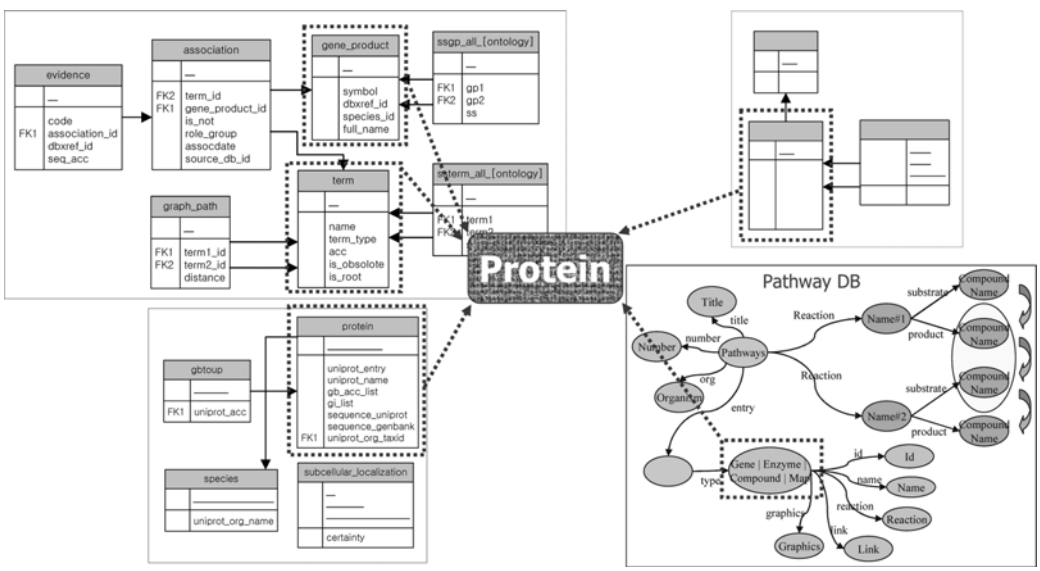


그림 3 원본 데이타베이스들의 스키마

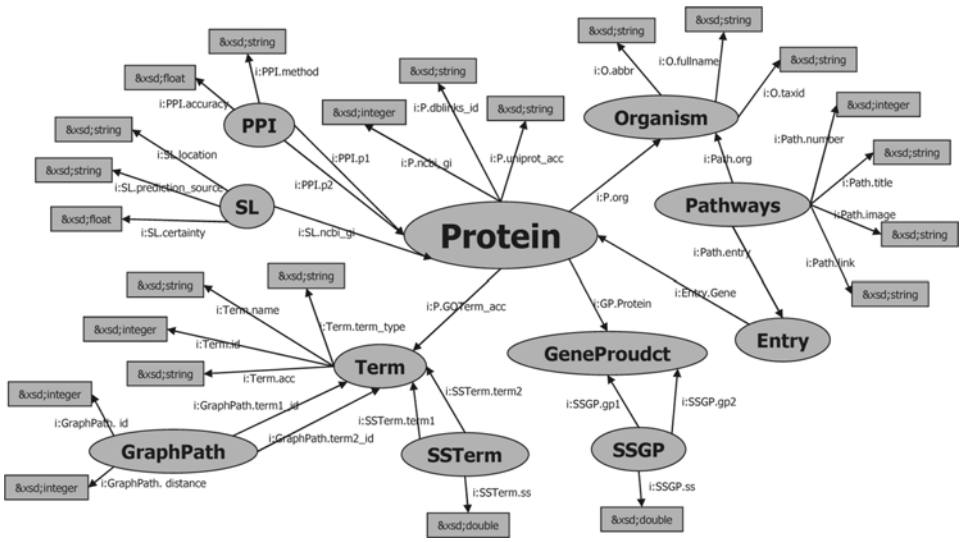


그림 4 통합 스키마

```
<!-- ==================== Classes ==================== -->
  <rdfs:Class rdf:ID="Organism"/>
  <rdfs:Class rdf:ID="Protein"/>
  <rdfs:Class rdf:ID="PPI"/>
  ...
<!-- ==================== Properties ==================== -->
  ...
  <rdf:Property rdf:ID="P.org">
    <rdfs:domain rdf:resource="#Protein"/>
    <rdfs:range rdf:resource="#Organism"/>
  </rdf:Property>
  <rdf:Property rdf:ID="P.ncbi_gi">
    <rdfs:domain rdf:resource="#Protein"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#INTEGER"/>
  </rdf:Property>
  ...
  <rdf:Property rdf:ID="PPI.p1">
      <rdfs:domain rdf:resource="#PPI"/>
      <rdfs:range rdf:resource="#Protein"/>
  </rdf:Property>
  <rdf:Property rdf:ID="PPI.p2">
    <rdfs:domain rdf:resource="#PPI"/>
    <rdfs:range rdf:resource="#Protein"/>
  </rdf:Property>
  <rdf:Property rdf:ID="PPI.method">
    <rdfs:domain rdf:resource="#PPI"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#STRING"/>
  </rdf:Property>
  <rdf:Property rdf:ID="PPI.accuracy">
    <rdfs:domain rdf:resource="#PPI"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#FLOAT"/>
  </rdf:Property>
  ...
```

그림 5 RDF(S)로 기술된 통합 스키마의 일부

3.3 표현적 모델: 래퍼와 통합 질의 처리

표현적 모델로서의 래퍼는 개념적 모델의 RDF 통합 스키마와 개별 원본 스키마들을 연결시켜주는 매개체 역할을 한다는 점에서 중요하다. 하나의 래퍼는 하나의 원본 데이타베이스와 통신한다. 래퍼는 통합 스키마가 필요로 하는 원본 데이타베이스의 정보를 파악하여 해당 데이타베이스로부터 적절한 정보를 가져온 후, 이를 저장소에 RDF 인스턴스(instance) 형태로 변환하여 저장한다. 각 래퍼가 RDF 인스턴스의 생성을 마치면, 이렇게 동적으로 생성된 인스턴스들을 대상으로 SeRQL[15] 질의를 처리할 수 있다.

래퍼가 원본 데이타베이스에 접근할 때는 RDF 매핑

정보 테이블(mapping information table)을 조회하여 원본 데이타베이스에 대한 정보를 얻는다. 매핑 정보 테이블에는 각 원본 데이타베이스의 형태(RDB/RDF), 주소(URL), 드라이버(driver), 이름, 접속 계정 정보 등의 기본 정보를 포함하여 그 데이타베이스의 어느 엔터티에 어떻게 접근하여 필요한 정보를 가져올 것인지가 RDF로 기술되어 있다. 그림 6에는 PPI DB에 접근하여 단백질 상호작용 예측값을 가져오기 위한 정보가 간략히 기술되어 있다. 래퍼는 원본 데이타베이스로부터 얻은 정보를 통합 스키마에 맞게 RDF 인스턴스로 변환하여 저장함으로써 자신의 임무를 마친다. 아울러 우리의 래퍼는 새로운 데이타베이스가 추가되어도 이에 유연하게 대처할 수 있다. 새로운 데이타베이스의 추가는 곧 새로운 의미의 추가를 의미하며, 그에 따라 통합 스키마를 수정해주고 그 데이타베이스와 통신할 수 있는 래퍼를 추가해주면 된다.

```
<i:srcInfo_RDB rdf:ID="srcInfo_RDB_2">
    <i:format>RDB</i:format>
    <i:dbUrl>brahma.snu.ac.kr</i:dbUrl>
    <i:dbDriver>com.mysql.jdbc.Driver</i:dbDriver>
    <i:dbName>testDB</i:dbName>
    <i:dbUser>testUser</i:dbUser>
    <i:dbPassword>testPW</i:dbPassword>
    <i:tableName>interaction</i:tableName>
    <i:columnName>pid1</i:columnName>
</i:srcInfo_RDB>
```

그림 6 RDF로 기술된 매핑 정보 테이블의 일부

그림 7은 사용자 통합 질의가 처리되는 전체 과정을 간략하게 표현한 것이다.

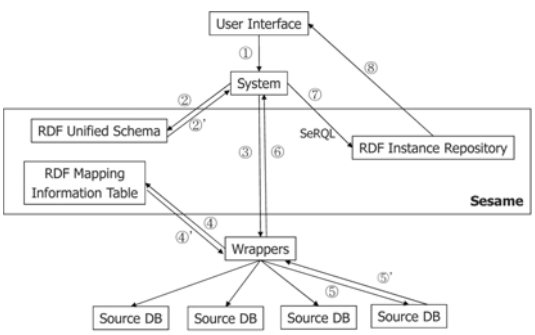


그림 7 통합 질의의 처리 과정

4. 결과분석 및 생물학적인 활용

4.1 결과분석

본 논문의 통합 시스템은 Sesame[14] 1.2를 이용하여 구현되었다. Sesame는 RDF 저장소로서 메인 메모리, 관계형 데이타베이스, 일반 파일을 사용할 수 있는데, 본 통합 시스템은 데이타의 크기가 크지 않은 특성상 메인 메모리 저장소를 채택하였다. 통합 레이어는 Java로 구현됐으며, 저장소에 접근하고 질의를 처리하기 위해 Sesame에서 제공하는 Java 기반의 Sail API를 이용하였다. 웹 서버(web server)로는 Tomcat을 이용하였으며, 웹 인터페이스(web interface)는 JSP로 씌어졌다.

본 논문에서 제안하는 통합 레이어는 단백질 정보의 대표라고 할 수 있는 4개의 데이타베이스들의 통합을 시도하였다. 표 2에서 정리된 바와 같이 이 4개의 데이타베이스들은 서로 다른 형태를 지니는 이종의 데이타베이스이다. 각각 RDB 혹은 일반 파일, RDF/XML 등의 형태로 데이타베이스에 저장되어 있으며, 각각 SQL과 SeRQL 등과 같은 다양한 질의어를 통해 질의된다. 데이타 통합에 관여하는 주요 데이타들은 작게는 수 MB의 작은 테이블에서부터 크게는 약 10GB에 이르기까지 그 크기가 다양하다. 각 데이타베이스의 정보들은 사용자의 통합질의가 요청되면 각각의 래퍼에 의해 실시간으로 질의되어 그 결과가 통합 레이어에 전달, 처리되기 때문에 통합 데이타베이스에는 정보가 데이타 웨어하우징 형태로 저장되지는 않는다.

4.2 활용 예 1: PPI 예측값의 검증

PPI DB는 두 단백질이 어느 정도의 연관성을 가지고 상호작용을 하는지에 대한 예측값을 제시해준다. 그러나 이는 어디까지나 예상되는 값으로서 불확실할 수 있기 때문에, 실제로 다른 데이타베이스의 단백질 정보를 비교함으로써 이 예측값의 타당성을 확인해보는 것은 생물학자들에게 있어 의미 있는 작업이라고 할 수 있다. 따라서 OASIS 환경에서는 PPI DB를 통해 서로 밀접한 연관을 지니고 상호작용하리라 예상되는 두 단백질이 실제로 같은 세포 내에 위치하는지, 그리고 같은 대사회로에 속하는지를 SL DB와 Pathway DB를 통해 확인할 수 있다. 그림 8과 그림 9는 이와 같은 절차를 보여주고 있다. 우선 Step 1에서 기준이 되는 단백질을 입력하면(Step 1에 대한 그림 생략), 그림 8과 같이 해당하는 단백질에 대한 정보가 화면에 표시된다. 화면 하단의 질의 입력 양식을 통해 통합 질의를 요구하면 그림 9와 같은 최종 결과를 얻을 수 있다. 결과 화면을 분석해보면, 기준 단백질인 16129833(NCBI[21]의 GI 값을 가리킴)은 대사회로 eco2030(KEGG[19] Pathway의 대사회로 ID를 가리킴)과 eco2031에 속하며, 세포 내 위치는 cytoplasmic임을 알 수 있다. 한편, 이 기준 단백질과 밀접하게 상호 작용을 할 것이라 예측된 단백질은 16129834와 16129835인데, 이들의 대사회로 역시 eco2030과 eco2031, 세포 내 위치도 cytoplasmic으로서 기준 단백질과 정확히 일치함을 보이고 있다. 이와 같이 어떤 단백질에 대한 PPI 예측값이 타당함을 Pathway DB와 SL DB와의 통합 환경 하에서 간단한 절차를 통

표 2 통합된 4개의 데이타베이스

DB	데이타 정보	데이타 포맷	DB 크기	통합에 관여하는 주요 데이타들 정보				질의문
				테이블	필드수	레코드수	사이즈	
GOA	Terms, gene products, annotations, etc.	RDB	134.3 MB	graph_path	4	1699591	27.6MB	SQL
				evidence	5	620591	17.4MB	
				association	7	603680	16.7MB	
				process_label	3	234791	15.5MB	
				gene_product	7	184895	10.5MB	
PPI	Interaction prediction value	RDB / flat files	9.3 GB	interaction	4	494792609	9.3GB	SQL
				protein	6	135506	4.3MB	
				organism	3	34	4.5KB	
Pathway	Pathways, entries, reactions, etc.	RDF/XML	458.3 MB	Total 1788 RDF files, size 87.5MB				SeRQL
SL	Localization prediction value	RDB	148.1 MB	subcellular_ localization	4	4416062	148.1MB	SQL

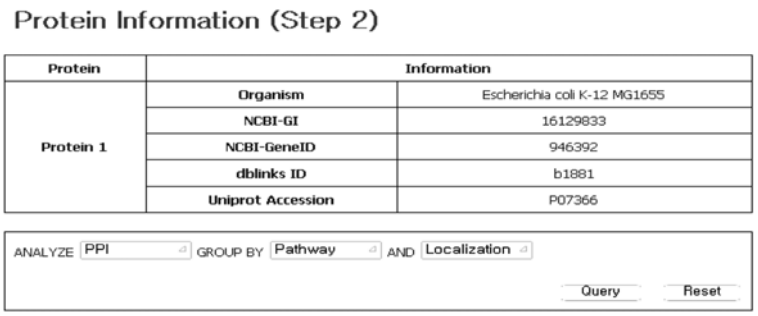


그림 8 기준이 되는 단백질의 선택

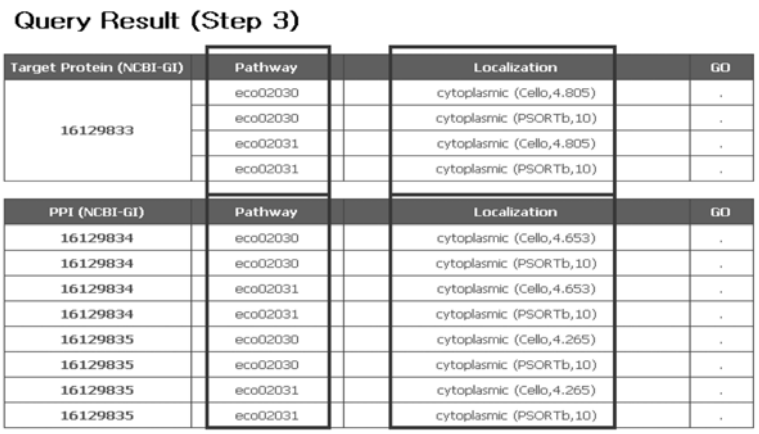


그림 9 통합 질의 결과 1

해 한눈에 확인할 수 있다.

4.3 활용 예 2: GO를 이용한 분석

KEGG Pathway[19]는 대사회로 지도(pathway map)를 이용하여 해당 유전자가 어떤 기능을 하는지를 직관적으로 표현해주고 있다. 그런데 하나의 유전자가 아닌 유전자들 간의 연관성을 알고 싶을 때 각각의 유전자가 서로 다른 대사회로 지도 상에 존재한다면 이들 간에 어떤 관계가 있는지 쉽게 파악하기가 어렵다. 사용자가 선택할 수 있는 방법은 브라우징을 통해 개별 유전자의 대사회로 정보를 확인한 후 이들간의 연관성을 직접 유추해내는 것이다. 하지만 통합 환경 내에서 Gene Ontology[6](이하 GO)를 이용하면 이 문제에 손쉽게 접근할 수 있다.

만약 우리가 관계를 알고 싶은 두 단백질이 GO 내에

서 어떤 term으로 기술되어 있는지에 대한 GOA 정보를 이용하여 다시 term들 사이의 관계를 알아낸다면, 두 단백질이 공유하는 특성을 확인할 수 있다. term들 사이의 관계라는 것은 GO가 그래프로 표현가능하기 때문에 그래프 내 위치에 따라 서로 is-a관계 또는 part-of 관계 등이 성립함을 가리킨다. 여기서 우리는 두 term의 공통된 특성을 찾기 위해 최소공통조상(least common ancestor, LCA)을 이용하였다. 최소공통조상은 한 트리에 속한 두 노드의 조상들을 찾아 올라갈 때 최초로 만나는 공통 조상 노드를 의미한다. 다만, GO는 트리가 아니라 DAG(Directed Acyclic Graph) 구조를 이루고 있기 때문에, 어떤 term과 최소공통조상 간의 거리(distance)가 하나 이상 존재할 수 있으며, 어떤 그래프 패스(graph path)를 따라 올라가느냐에 따라 최소공통조상이 하나 이상 존재할 수도 있다. 이에 우리는 두 term과 최소공통조상 사이의 거리의 합이 최소인 노드를 최소공통조상 노드로 선택하였다. 만약 두 term이 최소공통조상을 가지지 않는다면, 즉 루트 노드를 최소공통조상으로 가진다면 두 term은 아무런 관계를 가지고 있지 않다는 뜻이다.

그림 10은 이와 같은 방식으로 두 개의 단백질이 각각 매핑되어 있는 GO term들 간의 모든 조합에 대해 최소공통조상을 찾아낸 실험결과이다. 이 두 단백질은 KEGG Pathway 상에서 서로 다른 대사회로에 속하는 것으로 표현되어 있다. 첫 번째 행을 살펴보면, 단백질 1에 매핑되어 있는 GO:0008448(GO term을 가리키는 ID 체계)과 단백질 2에 매핑되어 있는 GO:0004565이 각각 거리 2, 거리 4의 조상으로서 GO:0016787을 가지며 이 노드가 두 term들의 최소공통조상이 된다. GO:0016787 밑의 부가정보를 보면 이 최소공통조상 노드는 molecular_function에 속하는 term으로서 이름이 hydrolase activity 임을 알 수 있다. 나머지 행들을 보아도 최소공통조상으로 대부분 hydrolase activity를 찾아내고 있는데, 우리는 이를 통해 이 두 단백질이 모두 대사회로 상에서 hydrolase activity라는 가수 분해효소 작용을 하고 있다고 결론지을 수 있다. 한편 최소공통조상으로 all을 찾아낸 행들도 있는데 이는 루트 노드로서 두 term이 서로 아무런 공통의 특성을 지니지 않는다는 뜻이다.

이 흥미로운 실험결과는 비단 대사회로 상에 존재하는 유전자에 대해서만 적용 가능한 것이 아니다. 단백질 정보라는 의미를 공유하고 있는 어떠한 데이타베이스도 그 단백질에 매핑되어 있는 GO term을 이용하여 다양한 의미의 정보를 GO로부터 얻을 수 있다. 가령, GOA DB에서는 주석 정보를 이용하여 term들 간의 의미적 유사성(semantic similarity) 값도 제공하고 있는데, PPI DB에서 서로 연관이 있다고 예측된 두 단백질의 GO term들이 실제로 어떤 의미적 유사성을 지니는지 GOA를 근거로 확인해 볼 수 있다. 이와 같이 본 통합 시스템을 통해 Is-a 관계 등과 같은 의미적인 질의도 훌륭히 처리해낼 수 있다.

5. 결 론

본 논문에서는 하나의 대상에 대해 다양한 형태와 의미를 지니는 여러 생물학 데이타베이스들이 RDF에 기반한 통합 레이어에 의해 개개의 데이타베이스로부터는 얻을 수 없는 의미 있는 정보를 이끌어 낼 수 있음을 보였다. 이 통합 레이어의 장점을 요약하면 다음과 같다.
(1) RDF의 장점들을 충분히 활용함으로써 형태론적인

GO Term Hierarchy

GO Term of Protein 1	↔ dist	Least Common Ancestor	↔ dist	GO Term of Protein 2
GO:0008448	2	GO:0016787 hydrolase activity (molecular_function)	4	GO:0004565
GO:0008448	2	GO:0016787 hydrolase activity (molecular_function)	0	GO:0016787
GO:0008448	2	GO:0016787 hydrolase activity (molecular_function)	2	GO:0004553
GO:0008448	2	GO:0016787 hydrolase activity (molecular_function)	1	GO:0016798
GO:0008448	5	all all (universal)	5	GO:0005975
GO:0008448	5	all all (universal)	4	GO:0009341
GO:0016787	0	GO:0016787 hydrolase activity (molecular_function)	4	GO:0004565
GO:0016787	0	GO:0016787 hydrolase activity (molecular_function)	0	GO:0016787
GO:0016787	0	GO:0016787	2	GO:0004553

그림 10 통합 질의 결과 2

통합뿐만 아니라, 전통적인 통합기법으로는 다루기 어려웠던 의미론적인 통합까지 이루어냈다.

(2) 데이타 통합의 대표적인 두 가지 방법인 데이타 웨어하우징 기법과, 요구 발생시 처리해주는 기법(on-demand retrieval)의 장점을 모두 취하였다. 사용자들은 마치 하나의 단일 데이타베이스만이 존재하는 것처럼 동적으로 생성된 RDF 인스턴스에 대해 질의할 수 있고, 원본 데이타베이스와의 동기화 문제도 고려할 필요가 없다.

(3) 원본 데이타베이스로부터 얻은 정보를 RDF 인스턴스로 손쉽게 변환할 수 있고, 새로운 데이타베이스가 추가 되어도 유연하게 대처할 수 있는 래퍼를 제안하였다.

생명 현상은 하나의 단편적인 정보만으로는 그 전체를 온전히 이해할 수 없으며, 필연적으로 다양하고도 복합적인 측면의 정보를 고려해야만 그에 대한 올바른 분석이 가능하다. 따라서 생물학 도메인에서의 데이타 통합의 활용 가능성은 매우 크다. 앞서 제시한 두 활용예제에서 보는 바와 같이, 다양한 데이타베이스들을 어떠한 의미를 기준으로 통합하였을 때 발생하는 장점은 단순히 데이타를 한 곳에 모아놓은 차원을 넘어서서 생물학자들이 기대하는, 혹은 기대하지 못한 새로운 의미의 정보를 발견할 수 있는 가능성을 열었다는 점에서 큰 의미가 있는 것이다.

또한, 웹 자원과 그들 사이의 관계를 모델링하기 위해 고안된 RDF는 생물학 데이타의 특성과 가장 잘 어울리는 기술 방법이기에 RDF를 통한 생물학 데이타 통합에 대한 연구는 충분히 가치 있는 작업이라고 할 수 있다. 본 통합 시스템에서 활용된 RDF의 장점들은 아직 미미하며 향후 활용 가능성은 더욱 크다. 그 중에서도 현재의 통합 스키마에 다수의 새로운 스키마가 추가될 때 RDF는 큰 역할을 할 수 있다. 현재 OASIS 시스템의 통합 스키마는 4개의 다소 단순한 스키마들을 대상으로 하였기 때문에 대부분 수작업으로 구축되었지만, 향후 시스템이 진화하여 다수의 스키마들이 추가될 때에는 자동화되고 지능화된 통합 기법이 요구될 것이다. 이때 바로 RDF가 지닌 시맨틱 웹의 장점을 활용할 수 있다. 만약 새로이 추가될 어떤 스키마가 있을 때 그 스키마에 대한 정보를 RDF로 잘 기술해놓는다면, 통합 시스템이 이 RDF 메타 데이타를 읽어 들여 분석한 후 자동으로 통합 스키마에 추가할 수 있는 것이다. 여기서 새 스키마를 RDF로 잘 기술해 놓는 것이 가장 중요한데, 이는 RDF의 풍부한 표현력으로 인해 가능하다. 이와 같이 RDF는 생물학 데이타의 통합에 있어서 여러 가지 어려운 문제를 해결해주는 중요한 역할을 할 수 있다.

본 논문에서 제안한 통합 레이어는 많은 연구 과제를 안고 있다. 무엇보다도 현재 OASIS 시스템에서 일종의 프로토타입으로 작동하고 있는 4개의 데이타베이스에 다수의 데이타베이스들이 추가될 때, 앞서 언급한 바와 같이 새로운 통합 환경에 얼마나 유연하게 대처할 수 있을지에 대한 고려가 필요하다. 새로운 스키마의 추가는 곧 새로운 의미의 추가를 뜻한다. 이에 따라 통합 스키마의 모습도 달라져야 할 것이다. 이 모든 과정에 효과적으로 대처할 수 있는 통합 시스템의 새로운 기능이 필요하리라 예상된다. 또한, 본 논문에서 제안하는 통합 스키마는 현재 스키마 레벨과 데이타 레벨(혹은 인스턴스 레벨)이 혼용되어 있다. 가령, 하나의 단백질 엔터티는 데이타 레벨이고 이와 맞물려 있는 GO[6]의 term들의 온톨로지는 스키마 레벨이지만, 통합 스키마에서는 이들 모두를 동일한 레벨로 표현하고 있다. 따라서 이를 더욱 정확하고 세련되게 표현할 수 있는 방법이 필요하다.

참 고 문 헌

[1] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. and Barabási, A.-L., "The large-scale organization of metabolic networks," Nature, Vol.407, pp. 651-654, 2000.

[2] Jeong, H., Mason, S. P., Barabási, A.-L. and Oltvai, Z. N., "Lethality and centrality in protein networks," Nature, Vol.411, pp. 41-42, 2001.

[3] Papin, J. A. and Palsson, B. O., "Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk," Journal of Theoretical Biology, Vol.227, pp. 283-297, 2004.

[4] O. Lassila, R. Swick, "Resource Description Framework (RDF) Model and Syntax Specification," W3C Recommendation, World Wide Web Consortium, 1999.

[5] The World Wide Web Consortium, http://www.w3.org/

[6] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, M., Davis, A., Dolinski, K., Dwight, S., Eppig, J. et al., "Gene Ontology: tool for the unification of biology," Nature Genetics, Vol.25, pp. 25-29, 2000.

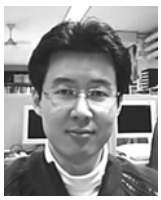
[7] Goldbeck, J., Fragoso, G., Hartel, F., Hendler, J., Parsia, B. and Oberthaler, J. "The national cancer institute's thesaurus and ontology," Journal of Web Semantics, Vol. 1, pp. 1-5, 2003.

[8] Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M. et al., "UniProt: the Universal Protein knowledgebase," Nucleic Acids Research, Vol.32, D115-D119, 2004.

[9] Dan Brickley, Ramanathan V. Guha, "RDF Vocabulary Description Language 1.0: RDF Schema," W3C Recommendation, World Wide Web Consortium, 2004.

[10] Thomas Hernandez and Subbarao Kambhampati, "Integration of Biological Sources: Current Systems and Challenges Ahead," ACM SIGMOD Record, Vol.33, Issue 3, pp. 51-60, 2004.
[11] Brenton Louie, Peter Mork, Fernando Martin-Sanchez, Alon Halevy, and Peter Tarczy-Hornoch, "Methodological Review: Data integration and genomic medicine," Journal of Biomedical Informatics, Vol.40, pp. 5-16, 2007.
[12] Kei-Hoi Cheung, Kevin Y. Yip, Andrew Smith, Remko deKnikker, Andy Masiar and Mark Gerstein, "YeastHub: a semantic web use case for integrating data in the life sciences domain," Bioinformatics, Vol.21, pp. 85-96, 2005.
[13] RDF Site Summary (RSS) 1.0, http://web.resource.org/rss/1.0/
[14] J. Broekstra, A. Kampman, F. Harmelen "Sesame: An Architecture for Storing and Querying RDF Data and Schema Information," International Semantic Web Conference, http://openrdf.org, 2002.
[15] Jeen Broekstra, Arjohn Kampman, "SeRQL: An RDF Query and Transformation Language," International Semantic Web Conference, 2004.
[16] Eric K. Neumann and Dennis Quan, "Biodash: A Semantic Web Dashboard for Drug Development," Pacific Symposium on Biocomputing, Vol.11, pp. 176-187, 2006.
[17] OASIS (Omics AnalySIS), http://idb.snu.ac.kr/
[18] Bowers P. M., Pellegrini M., Thompson M. J., Fierro J., Yeates T. O., Eisenberg D., "Prolinks: a database of protein functional linkages derived from coevolution," Genome Biology, Vol.5, No.5, R35, 2004.
[19] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, and Hidemasa Bono, "KEGG: Kyoto Encyclopedia of Genes and Genomes," Nucleic Acids Research, Vol.28, pp. 27-30, 2000.
[20] Pierre DÄonnes and Annette HÄoglund, "Predicting Protein Subcellular Localization: Past, Present, and Future," Genomics Proteomics Bioinformatics, Vol.2, pp. 209-215, 2004.
[21] NCBI (National Center for Biotechnology Information), http://www.ncbi.nih.gov/

이 강 표
2004년 연세대학교 컴퓨터과학과(학사) 2006년 서울대학교 컴퓨터공학부(석사) 2006년~현재 서울대학교 컴퓨터공학부 박사과정 재학중. 관심분야는 데이타베이스, 웹 2.0, 시맨틱웹



유 상 원
2000년 서울대학교 컴퓨터 공학과 졸업 2002년 서울대학교 컴퓨터공학부 석사 2002년~현재 서울대학교 컴퓨터공학부 박사과정. 관심분야는 데이타베이스, 생물정보학

김 형 주
1982년 서울대학교 전산학과(학사). 1985년 Univ. of Texas at Austin(석사) 1988년 Univ. of Texas at Austin(박사). 1988년~1988년 Univ. of Texas at Austin(Post-Doc). 1988~1990년 Georgia Institute of Technology(부교수). 1991년~현재 서울대학교 컴퓨터공학부 교수. 관심분야는 데이타베이스, XML, 생물정보학, 시멘틱웹, 웹 2.0