GO 기반의 SBML 문서 관리 및 질의 처리기

정승현° 정태성 김태경 김경란 조완섭
충북대학교 정보산업공학과, 경영정보학과
sane7142@nate.com , mispro97@naver.com ,
{tkkim, shira07, wscho}@cbnu.ac.kr

Gene Ontology based SBML Document Management and Query processing system

Seunghyun Jung°, Taesung Jung, Taekyung Kim, Kyoungran Kim, Wansup Cho
Chungbuk National University {Information Industry Engineering, MIS}

요 약

본 논문에서는 SBML문서를 효율적으로 저장관리 할 수 있는 Gene Ontology 기반의 SBML 문서관리 시스템을 제안한다. SBML은 시스템생물학에서 생화학적 네트워크 데이터의 교환 표준으로 연구 개발 되었으며, 다수의 생화학적 네트워크 데이터베이스들이 SBML을 이용하여 데이터를 제공해주고 있다. 이러한 SBML 문서를 통해 서로 다른 데이터베이스 또는 응용 프로그램간 정보를 교환으로 사용되고 있으며, 그 양 또한 급속하게 증가하고 있다. 따라서 본 논문에서는 이러한 대량의 SBML 문서를 효율적으로 저장, 검색 할 수 있는 문서관리시스템을 제안한다. 제안된 시스템은 OODB를 사용하여 효율적으로 SBML 문서를 저장관리하며, Gene Ontology를 기반으로 생화학적 용어의 모호성을 해결하고, SBML 문서간의 발생하는 데이터 중복을 제거하여 데이터의 품질을 제고하였다.

1. 서 론

지놈 프로젝트 이후 전산학을 이용한 생물학 분야인 생물 정보학과 그의 세부 분야인 시스템 생물학의 발달로 인하여 많은 양의 데이터와 복잡한 데이터가 나오게 되었다. 하지만 이들 데이터들은 서로 다른 형식으로 만들어지고 서로 다른 데이터베이스에서 관리 되고 있기 때문에 데이터의 활용도가 낮을 수밖에 없었다. 이러한 데이터 이질성의 문제점을 보안하기 위한 데이터 표준 모델과 데이터의 의미적인 복잡성을 해결하기 위한 연구가 진행 되고 있다.

특히, 바이오인포메틱스 분야에서 생화학적 네트워크 데이터에 대한 이해와 분석은 생명체를 이해하는데 매우 중요한 부분이다. 이러한 생화학적 네트워크 데이터를 표현 및 교환하기 위한 표준 모델로 SBML이 제안되었다[1]. SBML (Systems biology Markup Language)은 XML기반 언어로써 시스템 생물학에서 관련 모델들을 나타내는 표준(Standard)언어 중 하나이다.[1] SBML 데이터를 효과적으로 처리하는 방법에는 File System과 Relational Database, Object-Oriented Database, XML 전용 Database 와 같은 데이터를 다루는 여러 시스템에 의해서 처리 할 수 있다. 본 논문은 XML 데이터 처리에 있어서 객체지향 데이터베이스기반의 SBML처리 시스템을 제안 한다.

또한, 데이터의 의미적인 복잡성을 해결하기 위해서 유전자 온톨로지(Gene Ontology)를 이용하여 의미적 모호성을 해결하였다.

본 논문의 구성은 2절에서 제안된 시스템의 관련된 연구에 관하여 살펴보고, 3절에서는 본 논문에서 제안하고 있는 SBML 시스템의 구성과 기법에 대하여 살펴본다. 4절에서는 시스템을 이용한 구현과 활용에 대해서 알아보고, 5절에서 본 논문의 결론을 맺고자 한다.

2. 관련연구

2.1 SBML

SBML은 생화학적 반응에 대한 시스템을 네트워크로 묘사하고 있는 XML 기반 언어이다. SBML의 목적은 분산되어 있는 많은 데이터에 대해 표준 태그를 정의하고 있으며 데이터의 교환과 상호 운용적인 사용을 위해서 개발된 언어이다.[2][3][4][5] SBML은 생물 대사경로를 나타내는 cell signaling, metabolic, biochemical reactions, gene regulation 등의 여러 가지 데이터를 포함하고 있다.

SBML은 Level 1 스키마에서 시작 하여 현재 Level 2 스키마 버젼이 쓰이고 있다. 그리고 시뮬레이션 분석에 초점을 둔 Level 3 스키마가 이미 개발 중에 있다.

SBML은 현재 90가지 이상의 데이터베이스와 응용 프로그램에서 사용되어지고 있다. 이 시스템들은 대사경로를 시뮬레이션 하고, 시각화 툴을 통해 그래프로 보여질 수 있고, 데이터베이스에 저장 될 수 있다. SBML을 제공하는 Database로는 KEGG, BioCyc, Reactome 등 과 같은 것이 있다.

2.2 유전자 온톨로지

온톨로지는 특정 분양에서 어휘, 개념, 관계 등을 포함하는 특정 분야 지식 의미 모델이다. 즉, 온톨로지는 도메인 내의 지식을 개념화 하고 이를 명세 하는 것으로서 정의된다. 또한 어휘 사전의 역할 이외에 지식을 효과적으로 표현하기 위해 정보에 의미를 부여하고, 정보간의

관계를 설정한다. 이러한 장점을 이용하면 데이터통합을 위해서 이질적이고 분산되어 있는 데이터를 상호운용 적으로 수집할 수 있고, 데이터의 질적인 측면에서 정확도가 높은 데이터를 추출할 수 있다. 이러한 온톨로지의 장점으로 인해 본 연구에서 제안하는 시스템의 데이터의 질을 높일 수 있었다. [6][7][8]

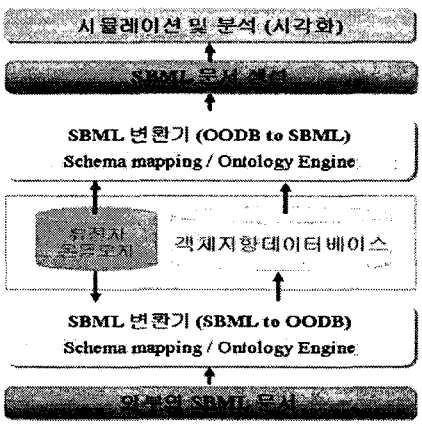
2.3 대사 경로 데이터베이스

시스템 생물학(Systems Biology) 분야에서 생화학 실험에 대한 데이터를 저장하고 관리하는 대표적인 데이터베이스는 KEGG, EcoCyc등을 들 수 있다. 이러한 데이터베이스는 화학 반응에 대한 대사경로와 관련된 데이터를 관리하기 위해 사용된다. 하지만 각 데이터베이스는 동일한 의미의 데이터를 저장하고 있으나 데이터의 형식이나 표준은 제 각기 다르므로 각 데이터베이스마다 그에 대응하는 어플리케이션을 새롭게 만들어야 하는 노력과 데이터의 교환에 문제점을 가지고 있다. JST ERATO Kitano Sysbiotic Systems Project 에서는 KEGG2SBML이라는 어플리케이션을 개발하였다. KEGG2SBML은 KEGG 데이터베이스에 저장 되어있는 정보를 SBML 문서로 변환하기 위한 어플리케이션이다. 하지만 KEGG는 관계형 데이터베이스를 이용하고 있다. 관계형 데이터베이스는 테이블의 참조 관계를 키로 연결하므로 테이블 수가 많아지고 복잡하다. 또한 복잡한 질의시 조인 비용이 높아진다. SBML을 활용하는 측면에서 관계형 데이터베이스의 문제점을 보완하기 위해서 우리는 객체지향 데이터베이스 기반의 시스템을 제안한다.

3. SBML Management System

3.1 시스템 아키텍쳐

본 논문에서 제안하는 SBML 시스템의 전체적인 구성을 살펴보면 [그림 1]과 같다.



[그림 1] SBML문서 관리 시스템 구성도

외부의 SBML문서를 객체지향 데이터베이스에 저장 할 수 있으며, 객체지향 데이터베이스에서 사용자가 검색한 결과를 SBML 문서로 변화하여 시뮬레이션 하거나 다른 여러 응용프로에 응용할 수 있는 시스템이다.

3.2 OODB & SBML 스키마의 관계

객체지향 데이터베이스를 이용하는 이유는 첫째로 객체지향 데이터베이스의 스키마처럼 SBML 스키마도 XML의 객체 지향적인 데이터 모델의 특성을 그대로 유지하게 되므로 SBML 스키마와 객체지향 데이터베이스 스키마와의 매핑이 쉬워진다. 매핑이 쉬워지므로 변환 과정이 보다 쉽고 간결해지는 장점을 얻을 수 있다. 두 번째로 객체지향 데이터베이스의 특징은 계층 구조와 상속관계 그리고 집합 값을 쉽게 표현해 줄 수 있다. 게다가 객체를 참조하는 구조로 되어 있어 질의가 간단해질 수 있는 장점을 얻을 수 있다. 셋째로 생물학적인 관점에서 SBML 기반의 스키마를 이용하므로 cell-signaling pathways, metabolic pathways, biochemical reactions, gene regulation등과 같은 시스템 생물학 분야를 통합적으로 관리 운용 할 수 있다.

따라서 본 연구에서 제안하는 시스템은 이를 이용함으로써 보다 쉽고 빠른 매핑 방법으로 SBML 문서를 객체지향 데이터베이스에 저장할 수 있는 장점이 있고, 반대로 객체지향 데이터베이스의 질의 결과를 SBML 문서로 변화하여 상호운용 적으로 데이터를 교환 할 수 있고, 시뮬레이션 및 분석 관련 도구에 활용 할 수 있다.

[표 1]에서는 SBML 표준 스키마의 UML 표기법의 예를 보여주고 있다. 반면 [표 2]에서는 실제 구현에 쓰인 객체지향 데이터베이스 스키마의 예를 보여주고 있다. 표와 같이 두 스키마가 전달하는 의미가 매우 유사함으로 서로 매핑이 간결하고 쉽다는 장점을 가질 수 있다.

Model
id : Oid {use="optional"}
name : string {use="optional"}
functionDefinition : FunctionDefinition [0...*]
unitDefinition : UnitDefinition [0...*]
compartment : Compartment [0...*]
species : Species [0...*]
parameter : Parameter [0...*]
rule : Rule [0...*]
reaction : Reaction [0...*]
event : Event [0...*]

[표 1] SBML 스키마의 UML 표기

Model
id : Oid
name : String
functionDefinition : FunctionDefinition {SET}
unitDefinition : UnitDefinition {SET}
compartment : Compartment {SET}
species : Species {SET}
parameter : Parameter {SET}
rule : Rule {SET}
reaction : Reaction {SET}
event : Event {SET}

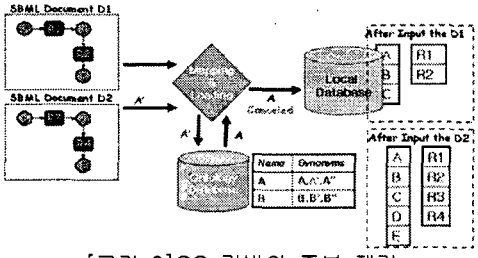
[표 2] OODB 스키마

3.3 시스템 처리 과정

3.3.1 GO(Gene Ontology) 기반의 중복제거

KEGG에서 제공하는 SBML문서는 각기 다른 169종에 대하여 총 12,122개의 Pathway model 정보를 표현 하고 있다. 제안된 SBML 문서 관리 시스템은 문서 하나가 들어올 때마다 SAX 파서를 이용하여 SBML 문서를 파싱하여 엘리먼트 이름, 타입 그리고 값을 가져온다. 그런

데 각 문서마다 동일한 의미의 용어가 다르게 표현되어 쓰일 수 있기 때문에 중복이 생길 수 있다. 그리고 하나의 데이터가 여러 문서에서 쓰이기 때문에 데이터베이스에 중복 저장될 수 있다. 이러한 중복 문제를 Gene Ontology와 자체 중복제거 모듈을 통해 해결 하였다. [그림 2]은 Gene Ontology를 기반으로 한 중복체크 모듈을 도식화한 것이다.



[그림 2]GO 기반의 중복 제거

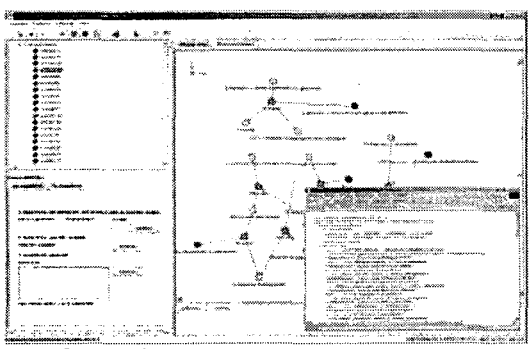
4. 시스템 구현 및 활용

4.1 OODB와 SBML 상호 연동

본 연구에서 핵심이 되는 기술은 KEGG나 Ecocyc 등에서 제공하는 SBML 문서나 외부에서 새롭게 작성된 문서들이 SBML 변환기를 통해 로컬로 구축된 객체지향 데이터베이스에 저장되는 것이며, 반대로는 로컬 객체지향 데이터베이스로부터 검색된 질의 결과를 SBML문서로 변화해 주는 과정이다. 3장에서와 같이 객체지향 데이터베이스의 스키마와 SBML 스키마가 쉬고 간결하게 매핑 되므로 변환 알고리즘이 간단하며, 시스템 성능을 향상 시킬 수 있다. 하지만 데이터의 변환과정이나, 검색 및 저장 시에 발생되는 중복이나 용어에 대한 모호성이 생길 수 있다. 이러한 문제점을 최소화하기 위해서 시스템 내부적인 중복 체크 알고리즘과 Gene Ontology를 사용하고 있다.

4.2 인터페이스

본 연구에서는 생물학 전문가나 일반 사용자들이 쉽게 원하는 정보를 검색하기 위한 인터페이스를 제안하고 있다. [그림 3]과 같이 사용자 인터페이스를 통해 다양한 형태로 검색을 하고, 그 결과를 시각화는 물론 SBML 문서 형태로 저장 및 교환할 수 있다.



[그림 3] SBML 문서 시각화 도구

4.3 SBML 문서 검증 및 시각화

본 연구에서 제안된 시스템으로 부터 나온 SBML 문서는 다양한 시뮬레이션이나 분석을 위한 어플리케이션의 시각화에 적용시킬 수 있다. SBML 문서를 이용하여 시뮬레이션이나 시각화를 수행하는 다양한 도구들이 제안되고 있으며 여기서는 이들을 생략한다. 본 연구에서는 완성된 SBML 문서를 시각화 어플리케이션인 JDesinger와 자체 Project인 Pathway Reconstruction tools를 이용하여서 테스트하고 있다.

5. 결론

최근 시스템 생물학에서 대두되는 표준화 문제는 SBML과 같은 XML을 기반으로 하는 몇 가지 언어(PSI MI, BIO PAX)를 통해 해결해 가고 있다. 이러한 언어를 통하여 다양한 시뮬레이션과 분석을 통해서 새로운 지식을 발견 할 수 있다. 결론적으로 이러한 XML 기반의 문서들을 효율적으로 처리 하는 것이 매우 중요하다. XML은 기존의 관계형 데이터베이스를 이용하여 데이터베이스에 저장하는 것 보다는 객체지향 데이터베이스를 사용하여 변환하는 것이 쉽게 처리 될 수 있다. 또한 온톨로지 기법을 적용함으로써 정보 검색에 있어 정확도를 높일 수 있다.

6. 참고문헌

[1] Michael Hucka, et al., "The ERATO Systems Biology Workbench : Architectural Evolution," 2001. 11.

[2] Andrew Finney "Systems Biology Markup Language(SBML) Level 2 . 2003. 6.

[3] M. Hucka, et al., "Evolving a Lingua Franca and Associated Software Infrastructure for Computational Systems Biology: The Systems Biology Markup Language (SBML) Project,"

[4] M. Hucka, et al., "The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models," 2003. 5.

[5] A. Finney "Systems Biology Markup Language: Level 2 and Beyond," *Trans*, 2003. 3.

[6] Jacob Kohler, et al., "SEMEDA: ontology based semantic integration of biological databases," 2003. 12.

[7] PG Baker, et al., ontology for bioinformatics applications," *Bioinformatics*, 1999. 6.

[8] http://www.geneontology.org