생화학적 네트워크 데이터의 효율적인 통합을 위한 시스템

정태성[1], 안명상[2], 조완섭[2]
{정보산업공학과[1], 경영정보학과[2]} 충북대학교
{mispro, epita55, wscho}@cbnu.ac.kr

A System To Integrate The Biochemical Network Data Efficiently

Taesung Jung[1], Myungsang Ahn[2], Wansup Cho[2]
{IIE[1], MIS} Chungbuk National University

요 약

유전자의 생물학적 기능을 밝히고 세포 내 상호작용을 이해하는 것은 post-genome era의 가장 중요한 작업 중 하나이다. 세포는 서로 다른 컴포넌트들의 상호작용에 의해 아주 복잡한 네트워크를 구성한다. 생화학적 네트워크에는 metabolic, regulatory, signal transduction과 같은 세포의 프로세스를 포함한다. 이러한 생화학적 네트워크들은 서로 다른 정보체계를 가지고 각기 다른 데이터베이스에 분산되어 저장관리 되고 있다. 따라서 생화학적 네트워크 데이터를 체계적으로 효율적으로 저장, 관리하기 위한 데이터베이스에 대한 필요성이 증대되고 있다. 본 논문에서는 기존의 생화학적 네트워크 데이터베이스의 장.단점을 분석하고 객체지향 방식에 입각한 새로운 생화학적 네트워크 데이터의 통합을 위한 시스템 모델을 제시한다. 제안된 시스템 모델은 생화학적 네트워크 데이터에 대한 생물학적 관계를 자연스럽게 표현할 수 있는 객체지향 모델을 사용하였다. 또한 생화학적 네트워크 모델을 묘사하기 위한 응용프로그램 사이의 데이터 교환의 표준언어인 SBML[2]스키마를 기반으로 하고 있다.

1. 서 론

유전자의 생물학적 기능을 밝히고 세포 내 상호작용을 이해하는 것은 post-genome era의 가장 중요한 작업 중 하나이다. 이런 목적을 위한 고전적인 방법은 먼저 어떤 형질의 원인이 되는 유전자를 발견하고 그 유전자의 구조를 밝히는 것이었다. 그러나 염기서열 해독의 자동화, 고속화, 대용량화가 급진전되고 컴퓨터를 사용한 정보처리 기술이 발전함에 따라 먼저 유전체의 서열분석을 통해 기능화 되고 있다[1]. 세포는 서로 다른 컴포넌트들의 상호작용에 의해 아주 복잡한 생화학적 네트워크를 구성한다. 생화학적 네트워크에는 대사경로, 조절, 신호전이 등과 같은 세포의 프로세스를 포함한다.

세포를 구성하는 수많은 상호작용을 표현하는 생화학적 네트워크를 구축하는 일은 매우 복잡한 작업이다. 따라서 생화학적 네트워크를 효율적으로 저장, 관리하기 위한 데이터베이스에 대한 필요성이 증대되고 있다. 이 논문에서는 기존의 생화학적 네트워크 데이터베이스의 장.단점을 분석하고 객체지향 방식에 입각한 새로운 생화학적 네트워크 통합 데이터베이스 모델을 제시한다. 제안된 데이터 모델은 모든 생화학적 네트워크 정보를 자연스럽게 표현할 수 있는 객체지향 모델을 사용하였다. 또한 생화학적 반응모델을 묘사하기 위한 응용프로그램 간 데이터 교환의 표준 언어인 SBML[2] 스키마를 기반으로 하고 있다.

본 논문의 구성을 보면 2장에서는 관련연구로 다양한 SBML과 생화학적 네트워크 데이터, 유전자 온톨로지 및 생물학 데이터베이스에 대해 살펴본다. 3장에서는 객체지향 데이터베이스와 SBML의 스키마의 매핑 및 시스템의 전체적인 구성에 대하여 설명하고 4장에서 결론을 내린다.

2. 관련연구

2.1 SBML과 생화학적 네트워크 데이터

SBML은 생화학적 반응에 대한 시스템을 네트워크로 묘사하고 있는 XML 기반 언어이다. 또한 SBML은 cell signaling pathways, metabolic pathways, biochemical reactions, gene regulation등 시스템 생물학과 관련된 여러 종류의 네트워크를 포괄하고 있다. SBML의 기본적인 목적은 분산되어 있는 많은 데이터에 대해 표준 태그를 정의하고 있으며, 데이터의 교환과 상호 운용적인 사용을 위해서 개발된 언어이다. 현재 SBML을 활용하기 위한 어플리케이션이 많이 개발되고 있다. 따라서 생물학 데이터베이스들을 통합하기 위해 구축되는 데이터 웨어하우스의 스키마를 SBML 스키마와 상호 운용될 수 있도록 구축함으로서 이 후의 상이한 데이터베이스나 응용프로그램간의 데이터 교환을 효율적으로 할 수 있는 장점을 가지게 된다.

현재 KEGG를 비롯한 여러 경로 데이타베이스들은 서로 다른 시스템간의 데이터교환의 목적으로 내부 데이터

를 SBML로 변환해주는 서비스를 실시하고 있다. 생화학적 경로에 대한 정보를 저장한 데이터베이스나 소프트웨어들은 네트워크 모델의 교환을 위하여 SBML을 이용하게 될 것이다. 따라서 이 논문에서 제시하는 데이터 웨어하우스는 SBML 스키마와 SBML문서의 장점을 최대한 이용한 방법을 제시한다.

2.2 온톨로지(Ontology)

온톨로지는 도메인 내에서 공유되는 데이터를 개념화한 형식적이고 명백한 규정이며, 이는 특정분야에서 사용되는 표준 어휘들의 집합이라고 할 수 있다. 즉, 온톨로지는 도메인 내의 지식을 개념화 하고 이를 명세화 하는 것으로서 정의된다. 또한 온톨로지는 어휘 사전의 역할 이외에 지식을 효과적으로 표현하기 위해 정보에 의미를 부여하고, 정보간의 관계를 설정한다. 따라서 온톨로지는 광범위한 도메인에 적용이 가능 하도록 표준을 제시함으로써 자원을 공유하고 재사용할 수 있으며 지식을 정확하게 표현할 수 있다.

온톨로지의 이러한 장점을 이용하면 데이터 통합을 위해서 이질적이고 분산되어 있는 데이터를 상호 운용적으로 수집할 수 있다. 데이터 통합에 있어 기존의 XML도 정보의 정확한 표현은 가능하지만, 의미 표현이 불가능하고 다양한 XML 형태의 표현으로 인하여 정보의 모호성을 증가시키는 단점을 지니고 있다. 이러한 데이터 의미에 대한 모호성을 줄이고 개념화된 정확한 정보를 추출하기 위한 목적에서 온톨로지가 개발되고 있다.

본 연구에서는 유전자 온톨로지(Gene Ontology:GO)[3]를 적용하여 정확도가 높은 데이터를 추출하고자 한다.

2.3 생물학 데이터베이스

시스템 생물학(Systems Biology)[4] 분야에서 생화학 실험에 대한 데이터를 저장하고 관리하는 대표적인 데이터베이스는 KEGG[5], EcoCyc[6]등을 들 수 있다. 이러한 데이터베이스는 화학 반응에 대한 대사경로와 관련된 데이터를 관리하기 위해 사용된다. 각 데이타베이스는 동일한 의미의 데이터를 저장하고 있으나 데이터의 형식이나 표현이 제각기 다르므로 각 데이터베이스마다 그에 대응하는 어플리케이션을 새롭게 만들어야 하는 노력과 데이터의 교환에 문제점을 가지고 있다. 최근 JST ERATO Kitano symbiotic Systems Project[7]에서는 KEGG(Kyoto Encyclopedia of Genes and Genomes) 데이터베이스에 저장된 정보를 SBML 문서로 변환하고, 변환된 SBML 문서를 활용하는 연구가 진행되고 있다.

3. 데이터 웨어하우스를 이용한 데이터 통합

3.1 객체 데이터베이스와 SBML 스키마의 매핑

데이터베이스 스키마와 SBML 스키마를 매핑하는 방법에 대하여 설명한다. 이러한 매핑 구조를 명확히 정의해 놓는 이유는 외부 데이터가 SBML변환기에 의해 SBML로 변환되고 변환된 SBML은 자연스럽게 데이터베이스로 매핑되기 때문이다. 기존에 연구에서는 관계형 데이터베이스를 이용하여 SBML을 변환하고 있다. 그러나 SBML은 객체지향 개념으로 설계되었기 때문에 관계 모델과 적합하지 않은 문제점이 발생한다.

본 연구에서 제안하는 SBML 변환기는 객체지향 데이터베스와 XML 스키마 기반인 SBML 스키마를 매핑한다. 객체지향 데이터베이스의 특징은 계층 구조와 상속관계 그리고 집합 값을 쉽게 표현해 줄 수 있다. 게다가 객체를 참조하는 구조로 되어 있어 질의문이 간단해 질수 있는 장점을 가진다. 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. [표 1]은 SBML 표준 스키마의 UML 표기법의 예를 보여주고 있다. 반면 [그림 1] 은 실제 구현에 쓰인 객체지향 데이터베이스 스키마의 예를 보여주고 있다.

3.2 시스템 구성

[그림 2]는 생물학 데이터 웨어하우스의 시스템 구성을 보여주고 있다. [그림 2]의 각각의 구성요소들은 아래와 같은 기능을 수행한다.

1) 데이터 추출(Wrapper)

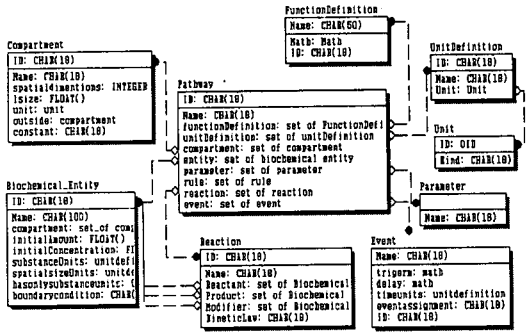
기존의 소스데이터는 플랫파일, 스프레드시트 및 xml문서 등의 다양하고 이질적인 포맷으로 구성 되어 있기 때문에 데이터를 추출하기 위해 각각의 소스데이터의 특성에 맞는 wrapper를 개발하고 이를 통해 자동적으로 데이터를 추출 저장하게 된다.

2)데이터 변환(SBML 변환기)

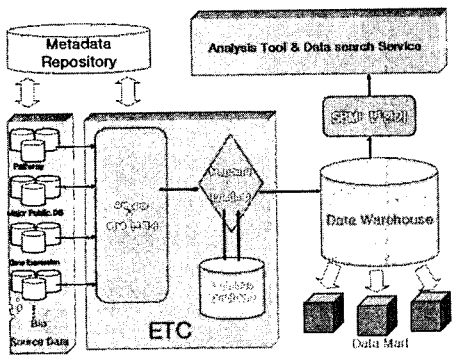
SBML 문서를 데이터 통합에 있어 가장 중요한 문제인 데이터의 중복제거 및 의미의 통합을 위해 Ontology Database와의 mapping을 통해 transformation 및 cleaning 과정을 거친다.

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



[그림 1] OODB 요약 스키마



[그림 2] 시스템 아키텍쳐

3)데이터 로딩(Data Loading and Mapping)

변환 및 정제 과정 이후의 통일된 의미와 형태의 데이터들은 로딩과정을 통해 데이터웨어하우스로 저장되며 목적에 따라 각각의 데이터 마트로 구축될 수 있다. 이때 객체데이터베이스를 사용하여 SBML문서 형태 그대로 데이터베이스 안에 논리적으로 저장될 수 있도록 한다.

4) 데이터 검색 및 분석(data search service)

사용자는 데이터 검색 서비스를 이용할 수 있으며, 검색 질의에 대한 결과는 SBML 변환기를 통해 SBML 문서가 자동적으로 생성된다. 시스템 구축을 위한 데이터베이스는 UniSQL을 사용 하였으며 데이터베이스의 연동을 위해서 UniSQL에서 제공하는 JDBC 드라이버를 이용 하였다. 제안된 시스템을 이용하여 현재 다수의 대사경로 데이터베이스와 유전자 정보 관련 데이터베이스로부터 데이터 웨어하우스를 구축하였다.

생물학 데이터 웨어하우스가 구축되면 OLAP (OnLine Analytical Processing)의 의사결정과 동등한 생물학적 의사결정을 할 수 있다. 의사결정자는 생물학 데이터에 대한 논리적이며 의미 있는 요약정보를 볼 수 있으며 필요할 때 drill-down과 같은 OLAP 연산을 이용하여 구체적인 데이터를 볼 수 있다.

4. 결론

최근 생물 데이타의 양이 기하급수적으로 증가되고 있으며 다양한 데이터를 효율적으로 처리하기 위한 데이터베이스들이 구축되었다. 그러나 생물 시스템의 분석은 특정 데이터에 의존하지 않으며 다양한 데이터들이 요구되는 통합 분석에 의해서 의미있는 결과를 얻을 수 있다. 따라서 이질적인 구조와 이질적인 포맷을 가진 데이터를 효율적으로 통합하기 위한 방법으로 생물 데이터 웨어하우스 구축 방법에 대하여 설명하였다. 제안된 시스템은 다음과 같은 특징을 갖는다. 첫째, 다양한 데이터를 표현하기 위해 객체지향 데이터 모델을 사용하였다. 둘째, 이질적인 데이터를 통합하기 위해 XML기반 시스템 생물학 표준인 SBML을 사용하였다. 셋째, 다양한 데이터의 의미를 통일시키고 중복 및 효율적인 통합을 위하여 온톨로지를 사용하였다.

향후 연구 과제로는 위와 같은 방법으로 구축된 데이터 웨어하우스에 생물시스템의 의사결정을 수행할 수 있는 OLAP 시스템에 대하여 연구해야할 것이다.

[참고문헌]

[1] Andrew Finney, Micael Hucka, "Systems Biology Markup Language(SBML) Level 2 : Structrues and Facilities for Model Definitions," 2003. 6

[2] David C., "Fallside(IBM) XML Schema Part 0 : Primer," 2001. 5.2

[3] Gene Ontology Available:http://www.geneontology.org/

[4] Stefan Hohmann, Jens Nielsen, Hiroaki Kitano, "Yeast Systems Biology - Concepts," 2004. 1

[5] Minoru Kanehisa, Susumu Goto, Shuichi Kawashima and Akihro Nakay, "The Kegg database at GenomeNet" 2001. 9.26

[6] Encyclopedia of *Escherichia coli* K12 Genes and Metabolism. Available:http://ecocyc.org/

[7] Akira Funahashi, Hiroaki Kitano, "Converting KEGG DB to SBML," ICSB2002, Stockholm, Dec. 12-15, 2002.