대사경로 데이터베이스 구축

안명상[O] ,정태성* ·조완섭** ·노동현***
[O]충북대경영정보학과, * 충북대정보산업공학과,**충북대경영정보학과, ,*** 충북대미생물학과
{epita55·mispro97,wscho,dhroh}@cbnu.ac.kr

On the Construction of an Object-Oriented Metabolic Pathway Database

Myoung-Sang Ahn[0], Tea-Sung Jung*,Wan-Sup Cho**, Dong-Hyun Rho***
[0]Dept. of MIS, *Dept. of IIE, **Dept. of MIS, *** Dept. of MicroBiology Chungbuk National University

요 약

유전자의 생물학적 기능을 밝히고 세포 내 상호작용을 이해하는 것은 post-genome era의 가장 중요한 작업 중 하나이다. 이러한 세포 내 상호작용은 복잡한 생화학적 네트워크를 형성하게 되며, 그 중 Metabolic pathway(대사 경로)는 생물 시스템을 이해하는데 가장 중요한 부분을 차지하게 된다. 대사 경로를 분석하기 위하여 분자의 기능 및 생화학적 프로세스에 대한 정보를 데이터베이스에 저장·관리해야하고, 사용자의 다양한 질의에 대하여 관련정보를 검색하여 GUI환경에서 제공해야 한다. 이 논문은 대사 경로 정보를 객체 데이타베이스 형태로 모델링하여 구축하고, 사용자가 관심있는 정보를 SBML형태로 제공하는 대사경로 데이타베이스의 설계 및 구현에 관해 다룬다.

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 기능을 추론한 후, 각 유전자의 기능을 확인하는 방식으로 변화되고 있다[1]. 세포는 서로 다른 컴포넌트들의 상호작용에 의해 아주 복잡한 생화학적 네트워크를 구성한다. 생화학적 네트워크에는 metabolic, regulatory, signal transduction과 같은 세포의 프로세스를 포함한다. 그 중 metabolic pathway(대사 경로)는 생물 시스템을 이해하는데 가장 중요한 부분을 차지하고 있다. 유전자들은 생화학적 반응을 촉매시키는 효소(Enzyme)에 대한 코드를 가지고 있으며, 이러한 생화학적 반응은 세포의 핵심기능을 구성하는 에너지와 여러 분자들을 생산하는 기능을 수행한다. 일련의 생화학적 반응들은 여러 효소와 중간 매개물과 연관되어 대사경로 지도를 이루게 된다. 대사 경로을 분석하기 위하여 분자의 기능 및 생화학적 프로세스에 대한 정보를 체계적으로 저장·관리해야 한다.

세포를 구성하는 수 많은 상호작용을 표현하는 대사경로 네트워크을 구축하는 일은 매우 복잡한 작업이다. 따라서 대사경로에 대한 데이터를 체계적이고 효율적으로 저장, 관리하기 위한 데이터베이스에 대한 필요성이 증대되고 있다. 이 논문에서는 기존의 대사경로 데이터베이스의 장·단점을 분석하고 객체지향 방식에 입각한 새로운 대사경로 데이터베이스 모델을 제시한다. 제안된 데이터 모델은 대사경로에 대한 생물학적 관계를 자연스럽게 표현할 수 있는 객체지향 모델을 사용하였다. 또한 생화학적 반응모델을 묘사하기 위한 응용프로그램간 데이터 교환의 표준언어인 SBML[2] 스키마를 기반으로 하고 있다.

이 논문의 구성은 다음과 같다. 2장에서 관련연구로서 대사경로와 SBML에 다하여 살펴보고 3장에서 객체지향 모델과 SBML을 통합한 데이터 모델을 제시한다. 마지막으로 4장에서 결론을 맺는다.

2.관련연구

2.1 대사경로 지도(metabolic pathway map)

대사경로 지도는 한 세포에서 상호 연결된 생화학적 반응의 네트워크이다. 단순한 박테리아에 대한 대사경로 네트워크 조차 매우 복잡하다. 따라서 생화학적 네트워크를 분석하기 위해 데이터베이스는 필수적인 도구이다. 이 데이터베이스는 기본적으로 대사경로에 대한 데이터 뿐만 아니라 반응에 참여하는 생화학적 객체에 대한 정보를 저장, 관리 해야 한다. 그림 1은 효소(enzyme)에 의해 촉매화

본 연구는 한국과학재단 특정기초연구사업 (R01-2003-000-11723-0)으로부터 지원을 받았음

되는 화학적 반응을 보여주고 있다.

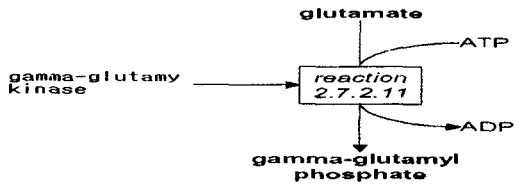

그림 1 효소촉매 반응

그림1에서 gamma-glutamyl kinase는 EC번호 2.7.2.11 반응을 촉매하는 효소이고 glutamate는 기질(substrate)로서 화학적 변화에 의해 gamma-glutamyl phosphate라는 화학적 생산물이 된다. 여기서 ATP와 ADP는 각각 cofactor-in, cofactor-out이다. 대사경로는 이러한 수많은 반응들이 서로 연관을 가지면서 구축된 복잡한 네트워크이다. 따라서 복잡한 대사경로 네트워크에 대한 데이터를 보다 체계적으로 저장,관리할 수 있는 데이터베이스 모델을 본 논문에서 제시한다.

2.2 대사경로 데이터베이스

대부분의 생물학 데이터베이스의 정보는 하나의 생물학적 엔트리를 기반으로 조직되어있다. 특히 유전자 데이터베이스들은 한 유전자마다 하나의 데이터베이스 엔트리로 표현되고, 유전자의 기능적 정보는 데이터베이스 엔트리의 설명필드에 free-text형태로 저장되어 있어 데이터의 중복이 발생하고 확장성에 대처하지 못하였다.

대사경로을 표현하기 위한 데이터베이스는 유전자 데이터베이스에 비해 좀더 데이터를 구조화시켜 표현하고 있지만 많은 한계를 가지고 있다. KEGG[3]는 완전한 유전체 서열을 기반으로 하는 분자의 상호작용을 분석하기 위한 웹기반 도구로서 광범위한 데이터를 저장하고 있다. KEGG의 데이타는 여러 데이터베이스로 분산되어 저장되어 있어 복잡한 링크를 구성하고 있다. 또한 대사경로에 참여하는 반응의 순서는 pathway 엔트리 자체에 저장되지 않으며, 그래픽 표현 시 반응의 순서를 고려하는 비 효율적인 구조를 가진다. EcoCyc[4]는 보다 정교한 계층적 데이터구조를 가지고 있다. aMAZE[5]은 생물학적 지식을 계층적으로 표현하기 위해 객체지향 모델을 이용하고 있으나 데이터베이스는 관계형 데이터베이스로 구 현되어 있어 객체지향 표현 모델과 관계형 구현모델간의 복잡한 매핑이 필요하다.

2.3 SBML(System Biology Markup Language)

SBML은 cell signaling pathways, metabolic pathways, biochemical reactions, gene regulation 등 생화학적인 반응 시스템들을 묘사하기 위한 XML 기반 언어이며 응용프로그램 간에 pathway 와 reaction model 에 관한 정보를 교환하기 위해서 사용된다. SBML 에서 화학적 반응은 여러 개념적 요소로 구성된다. 표 1 은 SBML 의 중요 구성요소를 간략하게 보여준다.

표 1 SBML 구성요소

구성요소	설명
Species	반응에 참여하는 화학적 객체
Reaction	생화학적 반응
Compartment	반응이 발생하는 위치에 대한 정보
Function	수학적 함수
Event	조건이 만족될 때 발생되는 이벤트
Unit	정량적 표현을 위한 unit 정의
Rule	모델에 대한 수학적 표현
Parameter	수학적 모델을 위한 변수 정의

현재 KEGG 를 비롯한 여러 대사경로 시스템들은 서로 다른 시스템간의 데이터교환의 목적으로 내부 데이터를 SBML 로 변환해주는 서비스를 실시하고 있다. 생화학적 경로에 대한 정보를 저장한 데이터베이스나 소프트웨어들은 네트워크 모델의 교환을 위하여 SBML 을 이용하게 될 것이다. 따라서 이 논문에서 제시하는 데이터베이스 모델은 SBML 스키마의 구조 및 장점을 최대한 이용한 모델을 제시한다.

3. 대사 경로 데이터베이스

3.1 대사경로 시스템 구성

그림 2은 대사경로 시스템의 전체적인 구조를 보여주고 있다. 이 시스템은 대사경로 데이타를 저장, 관리하기 위한 데이터베이스 부분과 사용자의 분석 질의와 시각화 서비스를 위한 서비스 제공 부분으로 크게 나누어진다.

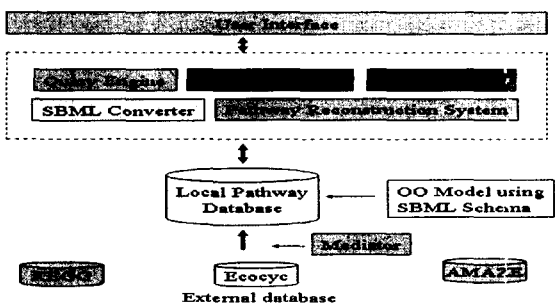

그림 2 시스템 구조

그림2에서 mediator는 이질적인 형식의 여러 데이터베이스로부터 데이터를 추출하여 로컬 테이타베이스 포멧에 맞게 변환하는 역할을 한다. 이렇게 구축된 로컬 데이터베이스를 이용하여 새롭게 밝혀진 유전자 서열 데이터에 대하여 pathway reconstruction system은 대사경로을 재구축하여 사용자에게 GUI환경에서 제공한다. 사용자 질의에 대해서는 질의의 결과를 sbml 형태로 보여주게 되는데 SBML Converter는 이런 기능을 수행한다.

3.2 대사경로 데이터베이스 모델

그림3는 제안된 데이터베이스 모델를 보여주고 있다. 이 모델은 객체지향 모델링 기법을 이용하여 대사경로를 자연스럽게 표현할 수 있으며, 효율적인 데이터 교환을 위하여 SBML 스키마를 데이터베이스 모델에 적용하였다. 아래는 제시된 모델의 특징 및 장점을 설명한다.

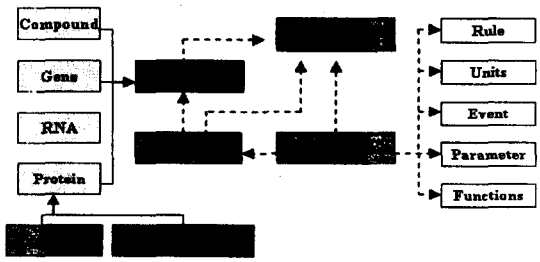

그림 3 데이타베이스 스키마

대사경로의 중요 구성요소인 생화학 반응 및 경로는 객체지향 DBMS에서 효율적으로 처리할 수 있는 복합타입(complex data type)이다. 따라서 생화학 경로는 자연스럽게 객체 그래프로 모델링 될 수 있으며, 전형적인 접근 방식인 그래프 순회방식을 사용할 수 있다. 게다가, 생화학 경로에 참여하는 객체들은 대부분 다-다 관계가 대부분이다. 이 경우 관계형 모델에서는 검색비용이 큰 조 인이 많이 발생하므로 DBMS의 성능이 저하된다. 이러한 이유로 인하여 객체지향 모델을 이용하였다.

그림 3의 데이터베이스 스키마는 크게 Bio Entity, Reaction, Pathway, Compartment 클래스로 구성된다. Bio Entity는 생화학 경로에 관여하는 생물학적 객체를 표현하는 클래스로서 각 객체들의 물리적 특성을 기술하는 속성정보를 가진다. Reaction은 기질(substrate)을 입력 받아 화학적 생산물을 출력하는 클래스로서 Bio Entity를 입력 받아 Bio Entity를 출력한다. 또한 Reaction은 반응의 촉매역할을 하는 효소, inhibitor, activator, cofactor-in, co-factor-out에 대한 속성정보를 가지고 있다. Pathway 클래스는 한 대사경로에 속하는 모든 reaction객체들을 참조하고 있는 클래스로서 완전한 대사경로를 표현한다. compartment클래스는 Bio Entity, Reaction, Pathway가 발생하는 위치에 대한 정보를 가진 클래스이다.

이 논문에서 제시한 모델은 SBML 스키마를 기반으로 하고 있다. 표1에서 species는 bio entity와 대응되는 개념이고, model는 pathway에 대응되는 개념이며 SBML의 다른 구성요소는 우리의 모델에 그대로 적용하였다. 따라서 저장된 데이터를 효율적으로 SBML문서로 변환할 수 있으며 역으로 SBML 데이터를 우리의 데이터베이스로 자연스럽게 입력시킬 수 있다.

3.3 구현

현재 개발중인 이 시스템은 SBML 스키마와 객체지향 모델을 기반으로 하는 대사경로 데이터베이스가 구축되어 있으며, 데이터베이스의 자료를 외부 시스템과의 교환을 위한 SBML converter가 구축되어 있다. 그 외 다른 구성요소는 향후 개발될 것이다.

4. 결론 및 향후계획

지금까지 대사경로 데이터베이스 시스템에 대한 구성요소와 데이터베이스 모델에 대하여 살펴보았다. 기술의 발전으로 인해 생화학 네트워크에 대한 데이터는 급속하게 증가할 것이다. 이러한 환경에서 객체지향 기법을 이용하면 대사경로를 자연스럽게 표현할 수 있으며, 새로운 지식을 유연하게 추가할 수 있다. 또한 다른 시스템간의 정보교환을 고려하여 XML을 적극 활용할 수 있는 모델이 필요하다. 우리는 전술한 바와 같이 필요에 부합하는 데이터베이스 모델을 제시하고 구현하였다. 향후 시스템 아키텍처에 제시된 다른 구성요소도 구현할 계획이다.

5. 참고문헌

[1] P.D Karp, " Pathway database:A case study in computational symbolic theories" ,*Science*, vol. 293, pp. 2040-2044, 2001
[2]M. Hucka, et al., " The System Biology Markup Language(SBML)" , *Bioinformatics*, vol. 19, pp.524-531,2002
[3] Kyoto Encyclopedia of Genes and Genomes (KEGG). [Online] Available : http://www.genome.ad.jp/kegg/
[4] Encyclopedia of *Escherichia coli* K12 Genes and Metabolism. Available:http://ecocyc.org/
[5] Jacques van Helden, et al., " representing and analyzing molecular and celluar function in the computer" , *Bio Chem*, vol. 381, pp.921-935