Biological Pathway 정보를 위한 바이오 온톨로지 기반의 XML DB

배은진[0] 박승수
이화여자대학교 컴퓨터학과
{jinnybae[0], sspark}@ewha.ac.kr

Bio Ontology-based XML DB for Biological Pathway

Eun Jin Bae[0] Seung Soo Park
Dept. of Computer Science & Engineering, Ewha Womans University

요 약

최근 바이오인포메틱스 분야는 IT의 발전과 더불어 급속한 발전을 이루어 왔다. 또한 기존의 분자생물학분야의 연구들을 컴퓨터의 첨단 기술을 이용하여 자동화, 전산화하는 것으로 생명정보의 데이터베이스 구축을 시작으로 서열정보의 주석처리 등이 발전되어 왔다. 또한 분자생물학의 오랜 연구결과들 중 수많은 문헌들을 토대로 한 분석인 텍스트마이닝, 정보추출은 전산학의 방법으로 데이터를 추출, 수집하여 분자생물학 데이터의 양적 팽창을 더욱 가속화하고 있다.

본 논문에서는 수집된 biological Pathway 정보의 효율적인 저장, 재사용 그리고 활용을 위해 바이오 온톨로지 기반의 XML DB를 설계하고자 한다. XML 스키마를 이용해 표현한 온톨로지를 기반으로 다양한 포맷의 생물학 데이터를 수집하고 수집된 데이터들로 XML DB를 구축한 후 Biological Pathway 정보 표현과 재사용, 나아가 생명정보학 응용분야에 활용하려 한다.

1. 서 론

생명정보학(Bioinformatics)의 데이터베이스는 분자생물학 데이터의 방대한 종류와 량에 있어 기하급수적으로 증가하는 정보를 RDB 또는 flat file로 저장하고 이를 관리, 활용하는 방안으로 진행되고 있다. 생명정보학의 데이터베이스는 오랜 연구 결과인 문헌들을 토대로 텍스트마이닝(Text Mining), 자연어처리(Natural Language Processing), 정보추출(Information Extraction) 등 전산학의 진보된 방법을 이용하여 바이오 데이터를 추출, 수집하여 양적 팽창을 더욱 가속화되고 있으며 대표적인 연구로 GENIA system이 있다.[1]

수많은 문헌으로부터 수집된 정보들 중 DNA, RNA, 단백질 등 데이터들은 그 자체만으로도 연구 대상이 되었으나 더 나아가 신체 조직 기관 안의 분자적 트랜젝션들과 프로세스들에 의해 기능들이 어떻게 수행되는지에 관한 Biological Pathway까지 포함하게 되었다.[2] Biological Pathway는 분자들의 상호작용에 의해 생체 시스템 상에서 기능들이 어떻게 수행되는지에 관한 연구 분야로 상호작용들을 그래프 형태의 지도를 사용하여 경로를 표현한다.[2]

지금까지 분자생물학 연구들은 각 기관이나 한 시스템에 국한된 연구들로서 다른 시스템간의 상호 호환성에 중요성을 두지 않고 진행되고 있다. 수많은 생물학 데이터들은 시스템 구현과정에 중복되어 수집, 가공되고 있으며, 시스템간의 효율적인 상호호환을 위한 데이터 관리(Data Management) 관점에 있어 데이터베이스 운영이 필요하겠다.

본 논문의 구성은 다음과 같다. 2장에서는 기존의 바이오 온톨로지의 특징과 응용 대해 알아보고 3장에서는 분자생물학 데이터의 특징과 XML과의 관계에 관해 알아보고, 4장에서는 본 논문에서 제안하는 XML DB의 구성에 관해 살펴보고, 5장에서는 결론과 향후 과제를 논의한다.

2. 관련연구

2.1 바이오 온톨로지(Bio Ontology)

온톨로지(Ontology)란 도메인에 국한된 지식의 세부명세로써 개념개층(Taxonomy)을 갖는 지식표현이다.[3] 바이오 온톨로지란 생물학 도메인에 관한 지식의 개념과 관계를 표현하기 위한 표현방법이다.[3]

2.1.1 온톨로지의 구성

온톨로지를 구성하는 주요한 구성요소는 다음과 같다.

- Concepts: 하나의 concept는 하나의 도메인 내의 개체 또는 'things'의 하나의 셋이나 클래스를 표현한다. primative concepts과 defined concepts가 있다.
- Relations: concept들 또는 하나의 concept의 속성

사이의 상호관계를 설명하며 크게 Taxonomies와 Associative relationships의 관계가 있다.

- Instances: 한 concept에 의해 표현되어진 'things' 이다.
- Axioms: 클래스 또는 인스턴스를 위한 values를 구성하기 위해 사용된다.

2.1.2 바이오 온톨로지의 응용

바이오 온톨로지는 응용 분야에 따라 community reference로 사용되며, database schema를 정의하거나 database annotation을 위한 공통의 용어(vocabulary)를 정의하기 위한 명세서로서의 온톨로지가 있으며, 정보들에 공통의 접근을 제공하기도 하고 데이터베이스에 있어 일관된 쿼리에 의한 온톨로지 기반의 검색을 지원하며, 데이터베이스 annotation과 기술문서의 이해를 위해 온톨로지를 이용하기도 한다. 기존에 개발되어 있는 바이오 온톨로지는 다음과 같다.

[표1] 기존의 바이오 온톨로지의 응용분야와 도메인

온톨로지	응용분야	도메인
RiboWeb	database schema	ribosome components, covalently bonded molecules, biological macromolecules, regions of molecules
EcoCyc	database schema	E.coli genes, metabolism, regulation, signal transduction and metabolic pathways
MBO	community reference	shellow
GO	controlled vocabulary for database annotation	drosophila, mouse and yeast gene function gene product function, process and cellular location and structure
TaO	common access ontology-based search	proteins, enzymes, motifs, secondary and tertiary structure, functions and processes, subcellular structure, and chemicals, including cofactors, the larger model includes nucleic acid and genes

이와 같이 분자생물학 분야에서 온톨로지의 사용은 최근 들어 활발히 사용되고 개발되고 있으나 이는 분자생물학 전체의 온톨로지가 아닌 부분적인 온톨로지로 개발되어 매우 상이하다는 특징이 있으며 온톨로지를 표현하기 위한 지식표현 방법도 표준화되어 정해져 있지도 않으며 기존의 온톨로지를 표현하기 위한 지식표현 방법으로는 Frame과 Description Logic을 이용하고 있다.

3. 분자생물학 데이터와 XML의 관계

3.1 분자생물학 데이터의 특징

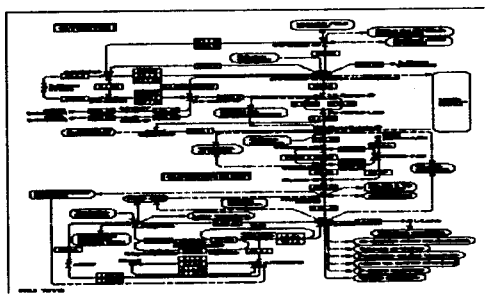
생물학 데이터(biochemical data)는 다른 데이터들과 달리 다음과 같은 특징을 갖는다.

- 생물학 데이터는 모델링 하기가 복잡하다: 수많은 관계를 나타내고는 많은 다른 타입의 데이터들이 있기 때문이다.
- 규칙적으로 새로운 타입의 데이터가 나타난다: 새로운 데이터 타입이 모델링 될 뿐만 아니라 기존의 개념(concepts) 사이에 새로운 관계가 나타나기도 한다. 즉 데이터를 모델링하고 통합하는 분석과정에서 새로운 데이터가 나타난다
- 원천 데이터(raw data)는 가공되지 않은 채 보존되어야 한다.
- 데이터의 매우 빈번한 업데이트가 발생한다.

3.1.1 Biological Pathway 정보

서론에서도 언급했듯이 분자생물학 정보는 DNA, RNA, 단백질 뿐만 아니라 분자들의 상호작용에 의해 생체 시스템상에서 기능들이 어떻게 수행되는지에 관한 연구[2]를 말하는 Biological Pathway에 관한 연구가 활발히 진행중이다.

이러한 biological pathway 정보는 생물학 데이터의 특징을 포함하면서 그들 간의 관계를 나타내므로 지식표현이 쉽지 않은 데이터이다. 기존의 정보들은 사람이 직접 손으로 그려 표현하고 있으며 이 과정을 자동화하려는 연구가 진행되고 있다.[2] [그림1]은 Biological Pathway의 대표적인 KEGG 시스템의 일부분이다.



[그림1] KEGG: Metabolic Pathway

3.2 생명정보학과 XML

XML(eXtensible Markup Language)는 구조화된 문서를 위한 표준을 나타내며, 인터넷을 통한 문서 교환을 위한 언어이다.[4] 분자생물학에서 XML은 다른 언어와 비교하여 데이터베이스와 다른 자원들 사이에 데이터의 상호 교환에 적합한 언어로서 위에 언급한 생물학 데이터를 수집, 가공, 저장, 표현하기에 적절한 언어이다.[4] 다음은 XML이 분자생물학에서 어떠한 장점을 보일 수 있는지를 나타낸다.

- XML은 매우 융통성이 있다: 새로운 요소나 속성을 추가하면서 DTD를 수정하는 것은 간단하며 이는 XML 데이터의 수정을 요구하지는 않는다. XML과 DTD 파일은 사람이 이해하기 쉽기 때문에 간단한 컴퓨터 기

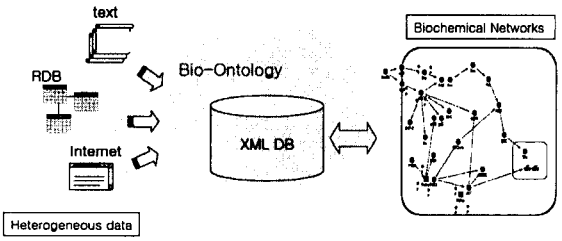
술을 가지고 쉽게 편집이 가능하다.

- XML은 인터넷 기반의 언어로서 연결 데이터를 위한 매우 풍부한 능력을 가지고 있다: 이것은 GenBank/EMBL/DDBJ 같은 데이터베이스 사이에서 상호 연결되어 레퍼런스로 대체되어 사용되어 진다.
- XML은 표준화된 명세서를 정의하기 위한 개방 framework을 제공한다: 이것은 표준화가 부족한 분자생물학 분야에 중요한 점이라고 할 수 있다.

4. 온톨로지 기반의 XML DB 시스템 설계

다양한 포맷과 기하급수적으로 증가하는 분자생물학 데이터를 저장하고 표현하기 위해서 그 데이터의 특성을 잘 반영할 수 있는 XML을 저장 포맷으로 사용하였다. 또한 기존의 Frame과 Description logic을 이용한 온톨로지를 XML 스키마로 재구성하여 변환한 후 이를 기반하여 XML DB를 설계하고자 한다.

특히 Biological Pathway의 하나인 Metabolic Pathway의 정보에 대하여 온톨로지를 구성하고자 한다. [그림2]는 Biological Pathway를 위한 바이오 온톨로지 기반의 XML DB를 구출하고 이를 활용하기 위한 전체 시나리오를 보여준다.



[그림2] Biological Pathway를 위한 바이오 온톨로지 기반의 XML DB

다양한 포맷의 데이터를 동일한 포맷의 flat file로 변환하여 수집하고 이때 기존의 지식표현을 이용한 온톨로지가 아닌 XML 스키마로 변환한 온톨로지에 기반하여 저장한다.

(1) 기존의 분자생물학의 부분적인 도메인을 표현하는 바이오 온톨로지중 Metabolic Pathway의 온톨로를 XML 스키마로 재구성한다.

(2) 재구성된 XML 스키마를 기본으로 기존의 수집된 데이터의 RDB의 릴레이션 스키마를 XML 스키마에 맵핑시킨다.

(3) XML 스키마에 따라 XML DB를 설계하고 데이터를 수집한다. 이를 이용 Metabolic Pathway를 표현한다.

(4) Biological Pathway표현에 사용된 데이터를 저장하고 재사용 한다.

```
.....
<xsd:element name="listOfCompartments">
 <xsd:complexType>
  <xsd:sequence>
   <xsd:element name="compartment"
    type="Compartment" maxOccurs="unbounded"/>
  </xsd:sequence>
 </xsd:complexType>
</xsd:element>
<xsd:element name="listOfSpecies">
 <xsd:complexType>
  <xsd:sequence>
   <xsd:element name="species" type="Species"
    maxOccurs="unbounded"/>
  </xsd:sequence>
 </xsd:complexType>
</xsd:element>
<xsd:element name="listOfReactions">
 <xsd:complexType>
  <xsd:sequence>
   <xsd:element name="reaction" type="Reaction"
    maxOccurs="unbounded"/>
  </xsd:sequence>
 </xsd:complexType>
</xsd:element>
.....
```

[그림3] XML 스키마의 element

```
.....
<xsd:complexType name="Specie">
 <xsd:complexContent>
  <xsd:extension base="SBase">
   <xsd:attribute name="name" type="SName"
    use="required"/>
   <xsd:attribute name="compartment"
    type="SName" use="required"/>
   <xsd:attribute name="initialAmount"
    type="xsd:double" use="required"/>
   <xsd:attribute name="units" type="SName"
    use="optional"/>
   <xsd:attribute name="boundaryCondition"
    type="xsd:boolean" use="default" value="false"/>
   <xsd:attribute name="charge" type="xsd:integer"
    use="optional"/>
  </xsd:extension>
 </xsd:complexContent>
</xsd:complexType>
.....
```

[그림4] XML 스키마의 attribute

[그림4]와 [그림5]의 XML 스키마의 element와 attribute를 표현한 일부분이다.

XML DB로 수집, 저장된 데이터들을 biochemical networks 정보를 그 도메인으로 정해 관련된 어플리케이션과 연동될 수 있다.

5. 결론 및 향후연구

생명정보학의 연구가 활발히 진행됨에 따라 지식표현의 하나인 온톨로지에 관해 살펴보고 기존의 생명정보학 분야에서 사용중인 바이오 온톨로지에 관해 알아보고 분자생물학 데이터의 특징과 그에 상응하는 XML의 특징을 살펴 보았으며 기존의 서열정보나 서열주석 데이터베이스와 달리 Biological Pathway 정보를 저장하고 표현하는 XML DB를 제시하고자 하였다. 이는 데이터의 중복된 수집과 가공, 산재 된 데이터들을 통합하고 생물학 데이터의 특성을 잘 반영할 수 있는 XML format으로 저장 관리하여 Biological Pathway 정보를 수집하고 저장, 표현, 재사용에 있어 중요한 역할을 할 것으로 본다.

향후 연구로는 본 논문에서 설계한 XML 스키마를 사용한 바이오 온톨로지 기반의 XML DB를 구축한 후 생명정보학에 활용하는 방안을 연구하자 한다.

6. 참고문헌

[1] GENIA. http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/

[2] KEGG. http://www.genome.ad.jp/kegg/

[3] Stevens R., Goble Carole and Sean Bechhofer, "Ontology-based Knowledge Representation for Bioinformatics", Briefings Bioinformatcs, 2001

[4] Achard F., Guy Vaysseix and Emmanuel Barillot, "XML, bioinformatics and data integration", Bioinformatics, Vol.17 no.2, pp.115-125, 2001

[5] Schulze-Kremer S., *Ontologies for Molecular Biology and Bioinformatics*, *In Silico* Biology 2, 0017, 2002