생물정보 데이터베이스 구축을 위한 XML 적용 기법

이범주(충북대학교 데이터베이스연구실) bjlee@dblab.chungbuk.ac.kr
박성희(충북대학교 데이터베이스연구실) hpark@dblab.chungbuk.ac.kr
류근호(충북대학교 데이터베이스연구실) khryu@dblab.chungbuk.ac.kr

요 약

최근 생물정보 분야에서는 웹상에서 단백질과 유전자의 서열정보 및 이와 관련된 실험과 참조정보를 다른 유전체 데이터베이스 시스템과 상호교환을 위한 표준 형식으로 XML을 이용하기 시작하였다. 더불어 웹에서 데이터전송을 위한 표준 형식인 XML을 생물정보응용 분야에서 이용하기 위한 BioML을 정의하였다. 그러나 BioML에서는 서열의 소스 및 참조정보, 갱신정보와 XML문서사이의 참조정보를 포함하지 않고 있다. 따라서 이 논문에서는 BioML에서 포함하지 않는 이러한 정보를 XLink와 XPointer를 이용하여 나타낼 수 있도록 BioML을 확장하고 BioML에 포함된 서열 정보에 대한 메타정보를 RDF를 적용하여 제시하였다. 이렇게 함으로써 이질적인 생명정보데이터베이스 시스템에서 서열에 대한 복잡한 링크 정보와 서열의 변경정보를 효율적으로 교환이 가능하다.

1. 서론

생물정보 분야에서 얻어지는 다양한 소스 데이터들은 이종의 포맷 및 동일한 염기 서열에 대한 정보조차 서로 다르게 표현된다. 따라서, 새로운 타입의 출현, 빈번한 업데이트 등과 같은 문제점을 안고 있다. XML에서는 이러한 문제점들에 대하여 효율적으로 구조화된 문서정의, 데이터 전송 및 상호참조를 위한 다양한 링크, 개방적인 표준 specification 정의 등을 제공하므로 생물정보 응용분야에 이용하기 위한 BioML언어(Bioinformatic Sequence Markup Language)를 정의하였다.

그러나, 이러한 BioML은 문서 자체에 대한 소스 및 메타 정보들, 문서들 상호간의 교환 및 참조정보를 포함하고 있지 않은 단점이 있다. 이 논문에서는 이러한 문제점들을 해결하기 위하여 소스 및 메타 정보들에 대하여 RDF를 적용하였고, 문서들 상호간의 참조를 위하여 Xlink, Xpointer 그리고 Xpath 적용을 제시한다.

2. Xpath를 이용한 이질적인 소스 데이터 링크

2.1 생물정보 데이터 특징과 XML.

다양한 소스 및 생물학 실험에서 생성되는 데이터들의 타입은 이종 형태의 매우 다양한 포맷으로 제작되며 규칙적인 업데이트와 빈번한 수정이 요구된다. 이러한 데이터들은 데이터 자체로 종료되지 않고, 새로운 데이터를 생성하기 위한 데이터로 사용가능하며 데이터 자체 분석만으로도 새로운 데이터를 생성한다. 따라서 위와 같은 데이터를 관리하기 위한 데이터베이스는 이종 형태의 데이터들을 상호 교환 및 검색할 수 있어야하며 빠른 스키마 변경을 지원할 수 있는 유연성을 가지고 있어야하며, 데이터 자체만을 저장하는 것이 아니라 데이터에 대한 메타 정보들까지 저장하여야한다.

XML은 위와 같은 특징을 가지고 있는 데이터들을 위해 만들어 졌으며 현재까지도 XSL,

XSLT, Xpath, Xlink와 같은 XML 지원 언어들을 생성하고 있다. XML데이터를 그래프 형태로 생성하기 위한 XSL은 변환언어(Transforming language : XSLT)와 서식언어(formatting language)로 분리되었으며, XML에서 매우 미흡하였던 Hyper-link를 제공하기 위해 Xpath, Xlink, Xpointer 언어들을 제공하였다.

2.2 BioML

단백질과 nucleotide sequence 정보에 대해 복잡한 주석처리 제공을 목적으로 하는 BioML은 XML 문서간의 상호 참조 및 데이터 교환과, 데이터 소스에 대한 메타 정보들을 나타내는데 미흡하다. 이러한 단점을 극복하고 BioML의 성능을 개선하기 위하여 문서간 상호 참조와 데이터 교환에 있어서는 Xlink와 Xpointer를 적용하였고, 데이터 소스에 대한 각종 메타 정보를 기술하는데 있어서 RDF를 적용하였다.

2.2.1 BioML에서 Xlink와 Xpointer를 이용한 원격, 이종문서간의 항해.

Xlink는 resource들 또는 resource의 일부사이에서 명확한 위치를 연결하는 것을 정의한다. 즉, 한 document에서 다른 document로 어떻게 연결되는가를 기술하고 있으며, 양방향 link가 가능하고, 많은 document와 document 집합들 사이도 연결이 가능할 뿐만 아니라 linkbase를 이용하여 link database에 연결을 저장하여 사용하는 것도 가능하다. 이러한 기능들은 Xlink에서 제공되는 attribute list들을 통하여 보다 유연하고 편리한 link를 제공한다.

Xpointer(XML Pointer Language)는 이종의 또는 타 문서 내부의 특정 위치를 참조하기 위해, 다시 말해 특정 문서의 일부분만을 선택할 때 유용하다. Xpointer는 정밀한 연결을 지원하며 특별한 tag 또는 mark(#)만의 삽입으로 연결이 가능하며, Xpath 노드에 추가적으로 포인트와 범위를 선택할 수 있게 한다. 이러한 포인트와 범위를 지원하기 위해 Xpointer는 노드 개념을 위치(Location) 개념으로 확장하였다. 모든 Xpointer의 위치 단계는 Axis, NodeTest, Predicate의 triple 표현법에 의거하여 매우 정확한 위치 경로를 생성할 수 있다.

3. BioML 메타데이터를 위한 RDF

범용 BioML 어플리케이션에서 이종의 그리고 거대한 양의 데이터를 처리하기 위해서는 표준화가 필요하다. 이러한 표준화 방법중 하나가 바로 RDF를 이용하는 것이다. BioML은 웹 상에서 컴퓨터가 이해할 수 있는 정보 교환을 목적으로 어플리케이션간에 데이터를 다루기 위한 표준화 방법이 필요하다. 즉, XML을 다루는 어플리케이션간에 상호 운용을 제공하기 위해 생물정보에 필요한 메타데이터를 처리하는 기반을 적용할 필요가 있다. 이러한 메타 정보는 웹 상의 생물정보 시스템에 대한 참조, 갱신정보 등의 설명을 교환하기 위해 필수적이며 XML 또는 비 XML resource를 포함하여 어느 타입이건 기술될 수 있다. 여기서는 BioML에 RDF(Resource Description Framework)를 적용하여 위와 같은 문제의 해결방안을 제시하였다.

4. 결론 및 향후 연구방향

생물정보학분야에서 생성되는 데이터들은 대용량 및 이질적인 포맷을 사용하는 문제점들을 안고 있다. XML은 이러한 문제점들에 대해 유연성 및 상호 연결 등에 있어서 매우 뛰어난 기능을 제공함에 따라서 생물정보학분야에서 현재 데이터 교환을 위해 BioML을 정의하였고 이에 대한 활용이 증가하고 있다. 또한, 이러한 BioML은 이기종 간의 데이터베이스 시스템 및 에플리케이션간의 상호운용성을 높인다. 따라서 이 논문에서는 이러한 XML을 기반으로 생물학 정보들의

기술 및 상호 응용을 목적으로 사용되는 BioML에 대하여 보다 뛰어난 기능들을 제공하기 위해 3.2절 BioML에서 Xlink와 Xpointer를 이용한 원격, 이기종간의 항해와 4절 BioML 메타데이터를 위한 RDF를 통하여 이질적인 데이터들과 이기종간의 데이터베이스 및 어플리케이션간의 우수한 상호 운용성을 제시하였다.