XML기반의 생물학적 서열 파일 포맷 변환 메카니즘

이영화[0] 박성희 김진수 류근호
충북대학교 데이터베이스 연구실
(lrh[0], shpark, khryu)@dblab.chungbuk.ac.kr, jskim@mail.paichai.co.kr

Biological sequence file format transfer based on xml technique

Rong Hua Li[0] Sung-Hee Park Jin Soo Kim Keon Ho Ryu
Database Laboratory, Chungbuk National University

요 약

현재 생명 정보는 웹 상에서 다양한 포맷으로 배포되고 있다. 이러한 생명 정보 분석을 위한 데이터베이스나 시스템마다 이질적인 포맷을 지원하고 있기 때문에 각 시스템에서 이용되는 포맷들간의 변환이 필요하다. 이러한 생명 정보의 포맷 변환은 1대1의 파서를 구현하여 진행하고 있으며 1:1 파서의 구현에는 많은 시간과 비용이 소모된다.

따라서, 이 논문에서는 생명 정보를 XML로 표현하고 이질적인 포맷간의 매핑 정보를 데이터베이스에 저장한다. 이러한 매핑 정보를 XML의 스타일 시트로 나타내어 최종적으로 원하는 포맷으로 변환한다. 이렇게 포맷 변환에 XML 기술을 이용함으로써 파서를 구현할 필요가 없이 매핑 정보를 스타일 시트로 기술하면 되기 때문에 구현이 용이하며, 원시 소스가 변경되었을 때 소스 전체를 수정할 필요가 없이 수정한 필드의 매핑 정보만 수정하고 그에 따라서 XSL을 수정하면 되기 때문에 원시 소스 변경의 영향을 많이 받지 않는다.

1. 서 론

1998년 W3C에서 XML을 차세대 인터넷상의 데이터 표현 및 교환의 표준으로 정하면서 특별한 수학적, 기술적, 과학적 정보를 표시하기 위하여 여러 분야에서 마크업 언어들이 표준화되었다. 이런 표준화된 언어에는 Mathematical Markup Language(MathML), Chemical Markup Language(CML), Geography Markup Language(GML) 등이 있다. 생명정보학 분야에서도 표준화 작업을 수행하고 있으나 아직 표준화가 되어있지 않다.

현재 대용량의 DNA염기와 단백질에 대한 서열, 구조, 실험 및 참조 정보에 대한 데이터베이스를 구축하여 웹 상에서 제공하고 있다. 그러나 웹으로 제공되는 이러한 유전체 관련 데이터들은 각기 데이터베이스마다 고유한 데이터 포맷을 가지고 있으며 서열 분석용 애플리케이션들도 서로 다른 포맷을 지원하고 있다. 예를 들어, NCBI[1]의 GenBank[2]는 데이터 포맷으로 ASN.1(Abstract Syntax Notation One)[3]을 사용하고 있으며 또 플랫 파일 형식으로 데이터를 배포한다. 또한 대표적인 서열 정렬 시스템인 BLAST에서는 FASTA포맷을 지원한다. 그런데 보통 생물학자들이 서열 데이터를 분석하고 연구할 때 어느 하나의 데이터베이스를 검색하고 애플리케이션을 사용하는 것이 아니라 여러 데이터베이스나 애플리케이션을 검색하거나 사용하기 때문에 매 번마다 시스템에서 지원하는 포맷으로 변환하여야 한다.

현재 이러한 포맷들간의 변환은 1:1의 파서를 구현하여 파싱하는 방법을 사용하고 있다. 예를 들어, NCBI에서는 그들의 데이터를 다른 포맷으로 표현할 때 표현하려는 포맷마다 하나의 파서를 구현하였다. 그러나 이러한 파서를 구현하려면 생물학자들이 두 개 포맷간의 매핑정보를 작성한 다음 전문적인 프로그래머가 특정 언어로 특정 플랫 폼에서 파서를 구현하여야 한다. 이런 1:1의 파서는 표현하려는 매개 포맷마다 하나의 파서를 구현해야 하며 또 플랫 폼이 바뀌었거나 소스 데이터의 포맷이 바뀌었을 때 거의 재활용을 못하고 매핑 정보도 다시 작성하고 그에 따라서 파서도 다시 구현하여야 한다. 때문에 서열 데이터를 쉽게 분석, 연구하기 위하여 용이하고 재활용 성이 좋은 포맷 변환 기술이 필요하다.

이 논문에서는 생명정보학의 서열 데이터를 GenBank의 BSML을 확장하여 XML로 표현하고, FASTA, EMBL, 또는 다른 포맷들간의 매핑 정보를 구성하여 데이터베이스에 저장한다. 이러한 매핑 정보를 사용자가 원하는 포맷의 XSL로 작성하여 XML문서에 적용함으로써 데이터 포맷을 변환하였다. XML문서에 표현하려는 포맷의 스타일 시트를 적용하면 표현하려는 포맷의 문서를 생성하였다. 이 메커니즘은 XML기술을 이용하여 이질 포맷간의 변환을 수행함으로써 서로 다른 생명 정보 데이터 분석 시스템간의 데이터 교환에 XML을 지원하는 플랫폼과 데이터베이스 기술을 활용할 수 있다.

2. 관련 연구

생명 정보 학분야 에서는 다양한 포맷으로 데이터를 표현하고 있다. 그 중에서 가장 널리 쓰이는 것이 플랫 파일 형태이다. 그러나 필드의 의미와 데이터 타입에 대한 일치되지 않은 점이 많이 있다. 또한 필드 자체가 애매 모호할 뿐만 아니라, 데이터베이스마다 제공하는 형식이 다르므로 전문가가 아니라면 매뉴얼을 참조해야만 내용을 이해할 수 있다. 따라서 인간의 상호작용이 없이는 프로그램에 의한 처리가 불가능하다. 이러한 플랫 파일은 EMBL[5], NCBI, DDBJ, SWISS-PROT과 PDB[6]등 데이터베이스에서 사용하고 있다.

생명 정보 학에서 가장 앞장 서고 있는 NCBI는 ASN.1 (Abstract Syntax Notation One)이라는 데이터 포맷을 주로 사용하고 있다. 이 포맷은 이진 데이터를 교환하는 수단이며 입력 데이터를 파싱하여 캡슐화하기 때문에 특정 응용 프로그램에 대한 파서가 없으면 사람이 식별할 수 없다. 또한 확장 성이 없으며, 질의를 사용할 수 없다는 단점을 내포하고 있다. []

XML(Extensible Markup Language)[4]은 1998년 W3C(World Wide Web Consortium)에 의해 웹을 기반으로 하는 구조화된 문서를 기술하고 표현하는 표준으로 지정되었다. XML은 모든 형태의 애플리케이션과 문서를 다룰 수 있는 수백 개의 새로운 마크 업 언어를 생성할 수 있게 해준다. 따라서, 수학 표현 식에서 MathML, 지리 정보 시스템 및 공간 메타 데이터 관리 시스템에서는 GML, 멀티미디어 응용 시스템에서는 SMIL, Web검색 엔진 등에서는 웹 문서에 대한 메타정보를 얻기 위해 RDF등을 이용한다[4]. XML 기술은 XSL, lt와 같은 xml문서를 다른 매체로 표현하기 위한 스타일 시트 언어, 문서들간의 링크 언어인 xlink와 문서 내부의 특정 엘리먼트나 엘리먼트의 내용의 위치를 나타내는 xpointer 등을 지원하고 있다[4]. 따라서 이질적인 생명 정보들간의 상호 참조 관계 표현이 용이하다.

생명 정보 데이터를 표현 하는 데는 BioML(Biopolymer Mark up Language)[7], BSML(Bioinformatics Sequence Markup Language)[8] 등 마크업 언어들이 있다. 그런데 서열 데이터의 복잡한 주석 처리 제공을 목적으로 하는 BioML이나 GenBank의 필드를 그대로 기술한 BSML이나 생명정보 데이터를 XML형식으로 표현하는 데만 초점을 맞췄고 XML기술이 지원하는 다른 언어들을 사용하여 문서간의 상호 참조 및 서로 다른 매체로의 표현에 대한 연구가 미흡하다.

3. 서열 데이터 포맷 변환 매카니즘

생명 정보 학의 데이터 포맷에는 GenBank, EMBL, PDB, FASTA, SCOP 등 이질적인 포맷들이 있다. 이러한 이질적인 포맷들간의 변환 매카니즘은 주로 세 가지의 모듈로 구성된다. 첫째, 이질적인 포맷으로 표현된 생명 정보 데이터를 XML문서로 기술한다. 둘째, 이러한 이질적인 포맷들간의 매핑 정보를 작성하고 그 매핑 정보를 이용하여 XSL로 기술하여 메타데이터로 데이터베이스에 저장한다. 셋째, 생명 정보를 기술한 XML문서에 데이터베이스에 저장된 XSL를 적용하여 생명 정보 학의 애플케이션에서 이용하는 포맷으로 변환한다.

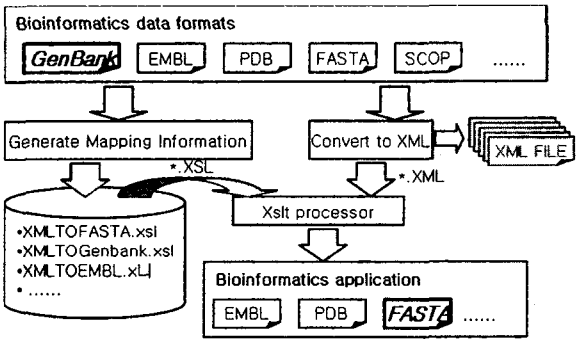


그림 1 XML를 적용한 서열 데이터 포맷 변환 매카니즘

이 논문에서는 GenBank의 마크업언어인 BSML을 XLink를 이용하여 확장하여 XML를 작성하고 GenBank, FASTA, EMBL 포맷들간의 매핑 정보를 작성하여 XSL로 기술하였다. 그리고 XML문서에 XMLTOFASTA.xsl을 적용하여 FASTA포맷으로 변환한다. FASTA, EMBL, GenBank포맷들간의 매핑정보를 작성하여 XSL로 기술한다. 이렇게 작성된 XSL을 XML 문서에 적용하여 FASTA 포맷으로 표현하였다.

3.1 생명 정보 서열 데이터의 XML화

핵산 서열을 가장 많이 포함하고 있는 NCBI의 GenBank에서는 이미 BSML이라는 마크업 언어를 정의하여 서열 데이터를 xml형식으로 나타내고 있다. 그러나 BSML은 그냥 플랫 파일의 내용을 그대로 기술하고 있을 뿐 XML문서들간의 상호 참조나 정보들간의 다양한 링크 정보에 대해서는 표현하지 못하고 있다. 생명 정보 서열 정보들간의 링크에는 버전 정보에 대한 링크와 서열과 관련된 문헌 정보들과의 링크가 있다. 버전 정보는 동일한 소스 대한 서열이지만 실험 방법이나 실험 환경의 변화에 따라 서열이 다를 수가 있다. 이때 이 서열을 원본 서열의 버전 서열이라고 한다. 서열 정보의 문헌 정보는 생물학자들이 주로 사용하는 정보이다. 문헌 정보에서 실험 방법이나 결과 같은 매우 중요한 정보들을 포함되고 있다. 하나의 서열은 여러 개의 버전 서열이 존재할 수 있으며 또 여러 문헌에서 이 서열에 대해 다룰 수 있다. 그림 2는 이러한 링크 정보를 나타내고 있다.

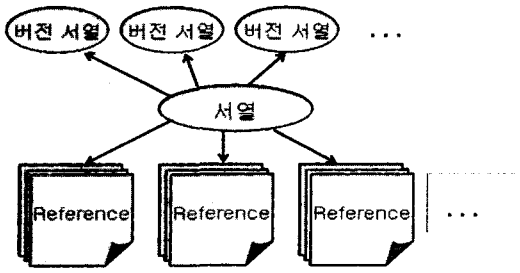


그림 2 서열 정보의 링크관계

이러한 정보를 XML의 링크 언어인 XLink를 이용하여 나타내면 양방향 링크나 1:n의 링크 같은 문서간의 다양한 링크 관계를 표현할 수 있다. 표 1는 그림 2의 서열 정보의 링크관계만 XLink를 이용하여 표현한 XML문서이다. 버전 정보를 나타내는 <vseq>와 문헌 정보를 나타내는 <Reference> 엘리먼트의 xlink:type속성을 "locator"로 줌으로써 버전 서열 정보와 문헌 정보의 리소스에 대한 링크를 나타내고 있다.

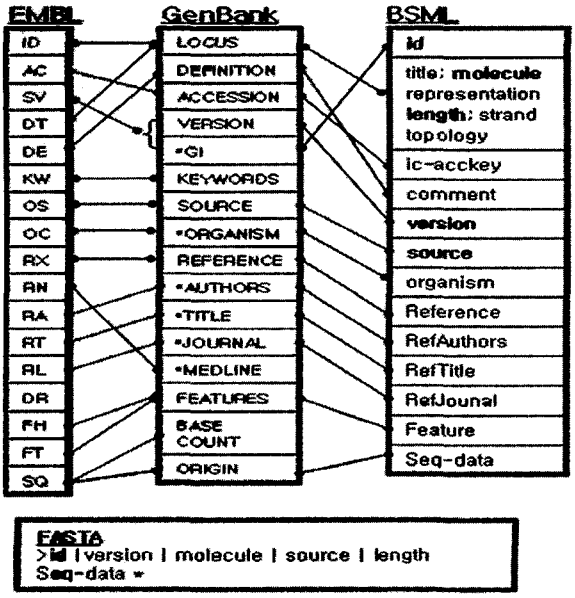
표 1 XLink를 이용한 링크 정보 표현

```
<sequence xlink:type="extended" >
   <id>44010</id> .....
   <vseq xlink:type = "locator" xlink:href = "44010/version/1.seq" xlink:label = "version1">
      <vid>1</vid> ..... </vseq>
   <vseq xlink:type = "locator" xlink:href = "44010/version/2.seq" xlink:label = "version2">
      <vid>2</vid> ..... </vseq>
   <reference xlink:type = "locator" xlink:href = "44010/reference/1.pdf" xlink:label = "ref1">
      <title>1</title>..... </vseq>
   <reference xlink:type = "locator" xlink:href = "44010/reference/2.pdf" xlink:label = "ref2">
      <title>2</title>..... </vseq>
   .....
</sequence>
```

3.2 이질적인 포맷간의 매핑정보 구성 및 스타일 시트 작성

표 3은 EMBL, GenBank의 플랫 파일의 필드, FASTA포맷과 BSML필드간의 매핑 정보이다. BSML은 GenBank의 필드를 기반으로 DTD가 정의 되었기 때문에 거의 1:1로 매핑된다. 그리고 EMBL과 GenBank는 통합된 데이터베이스를 사용하고 핵산 서열에 관한 데이터를 배포하고 있기 때문에 플랫 파일 필드들이 나타내는 값들이 비슷하지만 EMBL은 두 개의 자모로 필드

표 2 EMBL, GenBank, FASTA와 BSML 필드간의 매핑정보



들을 나타내고 있으며 GenBank는 단어들로 필드를 나타내고 있다. 대표적인 서열 유사성 검색 시스템 BLAST에서 사용하는 FASTA포맷은 >부터 시작하는 서열에 대한 주석 라인과 서열 정보 그리고 *로 서열의 끝을 나타낸다. 이 논문에서는 FASTA포맷의 주석 라인에는 XML문서의 id, version, comment 필드들의 값으로 표현하고 다음 라인에는 서열 데이터를 나타내고 있다. 위 표에서 1대 다로 매핑 되는 것은 하나의 필드가 여러 개의 정보를 포함하고 있는 경우 이다. 예를 들어, 는 GenBank의 LOCUS필드는 EMBL의 ID와 DT 두 개 필드의 값을 포함하고 있다. 이러한 매핑 정보를 이용하여 각 포맷의 XSL을 작성한다. 표 4는 매핑 정보를 이용하여 FASTA포맷으로 표현한 XSL문서이다.

표 3 FASTA 포맷의 xsl 소스

```
<?xml version="1.0" encoding="UTF-8"?>
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
xmlns:xs="http://www.w3.org/2001/XMLSchema">
  <xsl:template match="/">
    <html>
    <head/>
    <body>&gt;|<xsl:for-each select="seqset">
    <xsl:for-each select="sequence">
    <xsl:for-each select="@id"><xsl:value-of select="."/></xsl:for-each>
    |<xsl:for-each select="@versions"><xsl:value-of select="."/></xsl:for-each>
    |<xsl:for-each select="@molecule"><xsl:value-of select="."/></xsl:for-each>
    |<xsl:for-each select="@source"><xsl:value-of select="."/></xsl:for-each>
    |<xsl:for-each select="@length"><xsl:value-of select="."/></xsl:for-each>
    <br/><xsl:for-each select="seqdata"><xsl:apply-templates/></xsl:for-each>
    *</xsl:for-each>
    </xsl:for-each>
    </body>
    </html>
  </xsl:template>
</xsl:stylesheet>
```

3.3 XML문서에 스타일 시트 적용

3.2절에서 기술한 XSL문서를 3.1절에서 확장한 XML문서에 적용한다. 그러면 XML형식으로 표현된 서열 데이터를 EMBL이나 GenBank나 FASTA 포맷으로 표현할 수 있다.

4. 결론

현재 여러 분야 에서 XML를 이용한 표준화 작업이 이루어지고 있으며 웹 상에서 서로 다른 실험과 소스로부터 생성된 이질적인 데이터를 배포하고 있는 생명 정보 학에서도 데이터의 XML화가 필요하다.

이 논문에서는 NCBI의 GenBank 서열 데이터의 BSML문서를 XLink를 포함한 XML문서로 확장하고, GenBank의 데이터 포맷과 fasta, EMBL의 포맷간의 매핑정보를 구성하여 스타일 시트를 작성하고, 서열 데이터를 기술한 xml문서에 각각 스타일 시트를 적용하여 원하는 포맷으로 변환하여 표현하였다.

그러므로 xml기술을 생명 정보 데이터 포맷 변환에 사용함으로써 파서를 구현할 필요가 없이 매핑 정보를 스타일 시트로 기술하면 되기 때문에 구현이 용이하며; 원시 소스가 변경되었을 때 소스 전체를 수정할 필요가 없이 변경된 필드의 매핑 정보만 수정하고 그에 따라서 XSL을 수정하면 되기 때문에 원시 소스 변경의 영향을 많이 받지 않으며; 서로 다른 생명 정보 데이터 분석 시스템간의 데이터 교환에 XML을 지원하는 플랫폼과 데이터베이스 기술을 활용할 수 있다.

향후 연구로는 xml기술을 이용한 생명정보의 서열 데이터의 통합 시스템에 이러한 포맷 변환 매카니즘을 사용한다.

참고문헌

[1] David L. Wheeler, Deanna M. Church, Alex E. Lash, Detlef D. Leipe, Thomas L. Madden, Joan U. Pontius, Gregory D. Schuler, Lynn M. Schriml, Tatiana A. Tatusova, Lukas Wagner, and Barbara A. Rapp " Database resources of the National Center for Biotechnology Information: 2002 update" Nucl. Acids. Res. Vol:30. pp:13-16, 2002

[2] Dennis A. Benson, Ilene Karsch-Mizrachi, David J. Lipman, James Ostell, Barbara A. Rapp, and David L. Wheeler " GenBank" Nucl. Acids. Res. Vol: 30, pp:17-20, 2002.

[3]Stanley I. Letovsky, " Bioinformatics Database and Systems" Kluwer Academic Publishers, 2000.

[4] H.M.Deitel " XML How to program" Prentice-Hall, 2002.

[5] Guenter Stoesser, Wendy Baker, Alexandra van den Broek, Evelyn Camon, Maria Garcia-Pastor, Carola Kanz, Tamara Kulikova, Rasko Leinonen, Quan Lin, Vincent Lombard, Rodrigo Lopez, Nicole Redaschi, Peter Stoehr, Mary Ann Tuli, Katerina Tzouvara, and Robert Vaughan " The EMBL Nucleotide Sequence Database" Nucl. Acids. Res. Vol:30 pp:21-26. 2002.

[6] John Westbrook, Zukang Feng, Shri Jain, T. N. Bhat, Narmada Thanki, Veerasamy Ravichandran, Gary L. Gilliland, Wolfgang Bluhm, Helge Weissig, Douglas S. Greer, Philip E. Bourne, and Helen M. Berman "The Protein Data Bank: unifying the archive" Nucl. Acids. Res. Vol: 30, pp:245-248, 2002.

[7] http://www.bioml.com

[8] http://www.labbook.com

[9] 이영화, 박성희, 류근호, 홍순찬 "Xlink와 Xpointer를 적용한 유전체 데이터의 링크 설계" KISTI workshop pp: 27-42 2001

[10] Sung-Hee Park, Eun-Sun .Choi, Keun Ho Ryu, " Implementation of Implementation of Algebra and Data Model based on a Directed Graph for XML", Journal of Korean Information Processing Society vol. 8-D No. 6, 2001