실험실 레벨의 유전체 생물학 데이터베이스 관리시스템 구축

차효성[0] 정광수 박성희 류근호
충북대학교 데이터베이스 연구실
{kkido[0],ksjung, shpark, khryu}@dblab.chungbuk.ac.kr

Building a Biological Genomic Database Management System in Laboratory Level

Hyo Soung Cha[0], Kwang Su Jung, Sung Hee Park, Keun Ho Ryu
Database Laboratory, Chungbuk National University

요 약

대부분의 생물학 실험실에서는 스퀀싱 실험으로 얻어진 서열조각에 대해 어셈블리 과정을 통해 획득된 일치된 서열을 서열 실험파일 형태로 저장 한다. 이러한 서열 파일형태로 서열 데이터를 저장하면 사용자의 임의로 서열 정보 수정 및 서열 정보의 중복 등 서열 데이터에 대한 일관성 있고 무결성 있는 저장 관리가 어렵다. 또한 이질적 데이터 및 포맷을 통한 다양한 생물학적 분석이 요구된다.

따라서 이 논문에서는 시퀀싱을 통해 생성된 유전체 및 단백질 서열 데이터의 저장관리를 위해 서열 정보의 편집, 저장 및 검색과 서열 파일 포멧 변환을 수행하는 서열 정보관리 시스템의 구현을 목적으로 한다. 서열 저장시 서열 버전의 생성 및 검출을 위해 능동 데이터베이스의 트리거를 이용하여 시스템의 성능을 향상시킨다. 또한 서열정보 분석을 위해 이질적인 서열 포맷간의 포맷 변환은 서열 및 관련된 정보를 XML로 표현하고 포맷간의 매핑정보를 XML의 스타일 언어인 XSL을 적용하여 수행한다. 그러므로 원시 소스 변경시 영향을 적게 받으므로 이질적인 포맷간의 파서를 이용한 포맷 변환 보다 효율적이다.

1. 서론

HGP를 통하여 국내에서도 생명체의 유전체 및 단백질 서열에 대한 시퀀싱을 실행하고 서열데이터를 생산하고 있다. 주로 Template DNA를 단 한번의 single primer로 시퀀싱을 수행하는 단일 프리머 시퀀싱(Single-Primer Sequencing)과 PCR 산물을 이용한 시퀀싱을 서비스로 제공한다.

국내의 대부분 생물학 연구실에서는 시퀀싱된 서열 파일 정보 관리를 위한 소프트웨어가 존재하지 않아 파일 형태로 디스크에 저장해 두고 이용하는 상황이다. 따라서 사용자에 의해서 일부 서열 파일이 삭제되거나 변경된 내용이 반영되지 않을 수 있고 일관성 있게 서열 데이터가 관리되지 않는다. 그러므로 서열 데이터의 무결성과 일치성 있는 유지관리가 어렵고 이 때문에 장기적 연구에 활용하기 위한 서열 데이터 축적이 안된다.

국외에서는 NCBI[1], EMBL[2], DDBJ[3]등 인터넷을 통한 통합데이터베이스 검색 사이트로 생물 데이터 저장 및 검색 서비스와 분석 S/W를 제공한다. 또한 시퀀싱된 서열 데이터를 다른 연구자와 교환하거나 공개용 데이터베이스에 제출하기 위해서는 포맷변환을 자유롭게 할 수 있어야한다.

따라서 이 논문에서는 시퀀싱된 단백질 및 염기 서열 데이터의 효율적인 관리를 위해 서열 정보의 편집, 저장, 검색과 서열 파일 포맷 생성 및 변환을 수행하는 서열 정보관리 시스템의 설계와 구현을 목적으로 한다. 서열 정보 관리 시스템은 서열 파일로부터 읽어 전시하는 뷰어와 편집을 위한 서열 연산을 처리하는 서열 편집 관리기, 웹 데이터 표준 포맷으로 사용하는 XML을 서열 데이터 교환을 위한 공통 포맷으로 XML을 사용함으로써 호환성을 제공하는 서열 데이터 포맷 변환기, 서열 및 관련 정보를 데이터베이스에 저장할 수 있도록 비정형의 서열 정보를 저장하기 위한 데이터베이스 스키마를 설계하는 서열 저장관리기로 구성된다.

2. 관련연구

시퀀싱된 서열을 관리 및 분석하기 위한 소프트웨어로는 Staden Package[4]가 있다. Staden Package는 Cambridge 대학의 분자 생물학과의 의학 연구실(Medical Research Council Laboratory)에서 개발하였다. Staden Package는 시퀀싱된 실험 파일에 대한 뷰어인 Trev, 참조 데이터와 trace 데이터의 돌연변이 정보를 보여주는 trace_diff, 서열의 어셈블리 및 어셈블리 후 컨티그를 편집할 수 있는 GAP4, trace 데이터로부터 조각 어셈블리를 할 수 있도록 데이터를 준비하는 Pregap4, 어셈블리후 얻어지는 일치 서열을 분석할 있도록 서열 유사성 검색 및 연산을 제공하는 Spin으로 구성된다. Staden Package는 서열 파일에 포함된 서열 데이터를 분석 및 관리를 지원하지만 분석된 서열 데이터를 데이터베이스에 저장 및 검색을 지원하지 않는다. 서열 파일 기반의 분석프로그램으로 대량의 데이터를 처리할 수 없다.

NCBI[1]는 DNA 서열 데이터베이스인 GenBank, EST 서열 데이터베이스 dbEST, 단백질 분자 구조 모델링 데이터베이스 MMDB, 인간 유전자 카탈로그와 유전적 변이에 대한 데이터베이스 OMIN, 문헌정보 데이터베이스 PUBMED를 유지하고 있으면 이러한 이질적인 데이터베이스 사이에는 링크가 연결되어있다. 이러한 데이터베이

이 연구는2003년도 KISTEP의 특정연구개발과제의 지원으로 수행되었음.

스로부터 서열 및 유전자 데이터를 검색하기 위한 검색 시스템으로 Entrez를 이용하고 문헌 검색시스템으로 PudMed과 서열 유사성 검색 시스템으로 BLAST를 제공한다.

GenBank에서 데이터베이스의 서열 레코드를 식별하기 위해 Accession Number, Locus , GI(GenInfo Identifier)등 여러 가지 식별자를 사용한다. Accession Number는 Locus 이름보다 안정적이나 Accession으로 검색시 버전 서열에 대한 검색을 할 수 없다. 따라서 NCBI에서 유일하게 서열 데이터를 식별할 수 있는 GI(GenInfo Identifier)를 사용하게 되었다. 동일한 Accession Number를 갖는 서열에 대해서 서열이 변경될 때마다 변경된 서열에 새로운 GI를 할당한다. 이렇게 함으로써 서열의 검색 시 버전서열과 원본 서열에 모두에 대해서 검색이 가능하다. 이러한 식별자 정보는 NCBI의 공통포맷인 ASN.1[5]에 기술되어 있다.

3. 시스템 설계

그림 1 에서 서열 데이터 정보 관리 시스템은 서열 파일 포맷변환기, 서열 편집 관리기와 서열 저장 관리기 컴포넌트로 구성된다. 서열 포맷 변환기 컴포넌트에서는 디스크의 실험파일을 메모리로 로드하고 공통의 XML파일로 변환하기 위한 EXFileToXML모듈, 사용자가 원하는 파일 포맷으로 변환하기 위해 XML과 최종 파일간의 매핑정보를 데이터베이스에 유지하고 이러한 매핑 정보에 XSL을 적용하여 최종적으로 저장 및 전시할 수 있는 최종 포맷 생성 모듈로 이루어진다.

XML데 형태로 변환된 실험 파일은 서열 편집 관리기 모듈에서 서열 뷰어 모듈을 통해서 서열 정보를 전시하고 서열에 대한 연산을 여러 가지 연산을 서열 연산 처리기에서 수행한다. 이러한 서열을 저장하기 위해 서열 데이터 주석기에서 주석정보를 첨가할 수 있다.

서열 저장 관리기는 서열 버전관리, 서열저장기와 서열 정보검색기로 구성된다. 서열 버전 관리 모듈은 서열 및 주석 정보가 데이터베이스에 입력될 때 버전을 검사, 새로운 버전 생성 및 기존 버전 삭제를 관리한다. 서열 저장기는 새로운 서열 정보의 입력하고 서열 정보 검색기는 검색어를 입력받아 해당되는 서열정보를 전시한다.

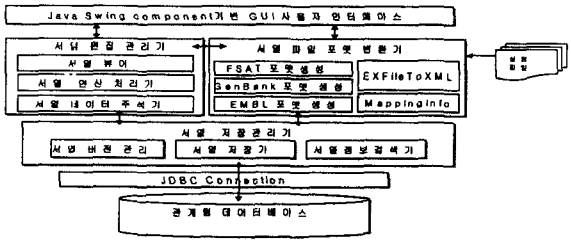

그림 1. 시스템 구조도

4. 서열 데이터 정보 관리 시스템

4.1 서열편집 관리

서열 뷰어 기능은 서열을 검색한 후 선택된 서열의 구성 베이스 함량과 시퀀스 데이터가 뷰잉 윈도우에 뿌려진다. 서열데이터 주석기는 편집중인 서열데이터에 필요한 주석정보를 추가한 후 데이터베이스에 단백질 식별자, 프로젝트 식별자, 서열타입, 소스생명체, 실험자 이름을 저장한다. 서열연산처리기는 서열을 구성하는 염기의 구성 비율을 계산하는 Base Composite, 서열의 특성 부분을 선택하여 새로운 엔트리로 구성하는 Set Range, DNA서열의 상보 서열 생성을 수행하는 Complement Sequence, 서열이 시작되는 위치를 새롭게 지정하기 위한 Rotate, DNA가 RNA로 전사될 때 DNA의 염기 중 티민(T)이 RNA에의 우라실(U)로 변경하는 Interconvert 연산들로 구성된다.

4.2 서열 파일 포맷 변환

XML을 이용한 이질적인 생명정보 포맷들간의 변환은 세 가지 모듈로 구성된다. 첫째, 어셈블리 과정을 통해 얻어낸 일치서열 및 관련 정보를 포함하는 실험파일로부터 서열 및 관련 데이터를 추출하여 XML문서로 기술하는 ExFileToXML 모듈과 둘째, 이질적 포맷들간의 매핑 정보를 작성하고 XSL로 기술하여 메타데이터로 데이터베이스에 저장하는 MappingInfo 모듈, 마지막으로 셋째, 해당 XSL 매핑정보를 적용하여 생명정보학의 애플케이션에서 지원하는 포맷으로 변환한다.

이 시스템에서는 실험실에서 추출된 서열 데이터에 초점을 맞춰서 실험파일을 표 1에서 정의 하였다.

표 1. 실험 파일 포맷 정의

Sequence ID	: 서열 식별 자
Name	: 핵산이나 단백질 이름
Version	: 서열 버전 정보
Molecule type	: 핵산 서열인가 아니면 단백질 서열인가
Sequence length	: 서열 길이
Date	: 서열 생성 날짜
Source	: 소스 생명체
DBref	: 참조한 데이터베이스에서의 식별자
Base count	: a, c, g, t의 개수 (단백질일 땐 생략)
Sequence	: 서열
Experimenter	: 시퀀싱 실험자

우리가 정의하는 FASTA포맷의 주석 라인에는 XML 문서의 id, version, molecule name, molecule type, organism, length 필드들의 값으로 표현하고 다음 라인에는 서열데이터를 나타낸다. 그림 2는 FASTA포맷의 예이다.

```
>1b9x_A|1|protein|Transduction|human|340
MSELDQLRQEAEQLKNQIRDARKACADATLSQITNNIDPVGRIQMRTRRTLRGHLAKI
YAMHWGTDSRLLLSASQDGKLIIWDSYTTNKVHAIPLRSSWVMTCAYAPSGNYVAC
LDNICSIYVMSLSLAPDTRLFVSGACDASAKLWDVREGMCRQTFTGHESDINAICFF
NGNAFATGSDDATCRLFDLRADQELMTYSHDNIICGITSVSFSKSGRLLLAGYDDFNC
NVWDALKADRAGVLAGHDNRVSCLGVTDDGMAVATGSWDSFLKIWN*
```

그림 2. FASTA 포맷 예제

표 1에 정의한 실험 파일 포맷을 이용하여 그림 3과 같이 XML파일을 작성하였다. XML문서에서 Seq-set 엘리먼트는 서열의 집합을 집합을 포함하는 루트엘리먼트이다. Sequence 엘리먼트의 어트리뷰트로는 식별자, 분자타입, 서열 길이, 분자 이름, 생성 날짜가 해당된다.

attribute 엘리먼트는 지속적으로 추가되며 현재는 소스 생명체, 버전 서열의 수를 포함한다. Version 엘리먼트는 버전 정보를 Feature-tables 엘리먼트는 레퍼런스 정보와 Feature 속성을 표현한다.

```
<?xml version="1.0" encoding="UTF-8" ?>
<seq-set>
 <sequence id="AB003468" molecule="dna" length="5350" name="cloning vector pAP3neo DNA"
  date="13-MAR-1999">
  <attribute name="organism" content="cloning vector pAP3neo" />
  <attribute name="versions" content= 1" />
  <seqdata type= original >gtagaaagcaccgacaatactcctggcatggcgttaaagctcacaggat</seqdata>
  <seqdata type= complement >atcctgtgagctttaacgcccatgccaggagtattgtcggtgcttctac</seqdata>
  <version id="AB003468.1  versionNum=1 >
   <vseqdata>ggctgtgtgcacgaaccccccgttcagcccgaccgctgcg</vseq-data>
   <diff><point>( 5,a,t)</point><point>(10,a,t)</point></diff>
  </version>
  <feature-tables>
    <Feature-table>
      <Reference dbxref="85176928"  title="1 (bases 1 to 4723)">
       <RefAuthors> Ebina,Y., Ellis,L., Jarnagin,K., Edery,M., Graf,L., Clauser,E.,
         Maslarz,F., Kan,Y.W., Goldfine,I.D., Roth,R.A. and Rutter,W.J.</RefAuthors>
       <RefTitle> The human insulin receptor cDNA: the structural basis for

hormone-activated transmembrane signaling </RefTitle>
        <RefJournal> Cell 40 (4), 747-758 (1985)</RefJournal>
      </Reference>
      <Feature id="FTR1"class="SOURCE     value-type="source" display-auto="0
        title="source"></Feature>
    </feature-table>
  </features-tables>
 </sequence>
</seq-set>
```

그림 3. XML 파일 예제

GenBank의 데이터 파일에서부터 서열 데이터와 그와 관련된 정보를 추출하여 XML문서로 작성하고 그림4와 같이 GenBank, FASTA, EMBL포맷들간의 매핑정보를 작성하여 XSL로 기술하고 XMLTOFASTA.xsl을 적용하여 FASTA포맷으로 변환과정을 보여준다. 또한 매핑정보를 이용하여 각 포맷의 XSL을 작성하고 웹브라우저를 통하여 결과를 확인 할 수 있다.

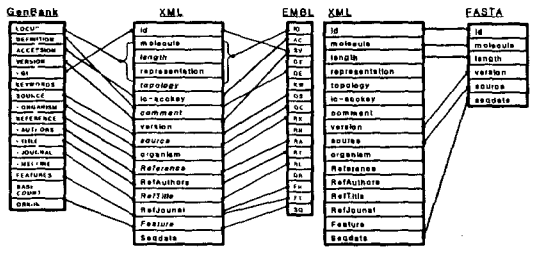

그림 4. DNA 서열 포맷간의 매핑 정보

4.3 서열 저장관리

서열 저장 관리 데이터베이스는 시퀀싱 실험을 통해 어셈블리 과정을 거친 염기 및 단백질 서열 데이터 및 관련 정보를 대상으로 데이터베이스화되었다. 서열데이터는 고유의 서열 식별자를 가지고 있으며 이 식별자에 의해서 구별된다.

서열 및 관련 정보를 저장하기 위한 테이블로는 Annotation, DerivedSeq, VersionSeq, PSequence 와 DSequence로 구성된다. 서열 데이터베이스의 관계형 스키마는 그림 5과 같다.

4.4 Trigger를 적용한 서열 버전 관리기

새로운 서열이 입력될 때 트리거를 이용하여 서열 테이블에 동일한 서열 식별자가 있는지 검사한다. 이러한 트리거 알고리즘을 간단히 표현하면 다음과 같다.

그림 6의 트리거를 이용하여 서열의 버전 가능성 여부를 가려 테이블에 새로운 서열을 입력한다.

Project	*Pid, Pname, Pdate, Puser*	시퀀싱 실험정보
Annotation	*ASid, Apid, Aog, Aname, Amachine, Adc*	부가서열 주석정보
DSequence	Sid, SPid, Slength, SnumVerSeq, Sdate, Sseq, SnumA, SnumT, SnumC, SnumG	염기서열
PSequence	Sid, SPid, *Slength, SnumVerSeq*, Sdate, Sseq	단백질서열
DerivedSeq	DSid, DPid, Dop, Dseq, *Dlength, Dtype*, Ddate	유도서열
VersionSeq	VSid, Vid, Vseq, Vlength, Vdate, Vdiffer	버전서열

그림 5. 서열 데이터베이스에 대한 관계형 스키마

```
INPUT : Sseq(새로운서열), SID(서열 ID)
1) 서열 ID를 이용 DB내의 서열 검색
2) SELECT COUNT(*) INTO Duple FROM Sequence
   WHERE Sid=newrow.Sid AND Sseq=newrow.Sseq;
 - IF Duple=0 THEN 중복서열 검사
 - IF Seqtrue=0 THEN 서열테이블에 저장
     ELSE COUNT(*) INTO Vertrue FROM VersionSeq
     IF Firstver=0 THEN INSERT VersionSeq
 - IF Verture≠0 THEN RAISE_APPLICATION_ERROR
 - IF Verture=0 THEN 버전 테이블저장
```

그림 6. 버전관리를 위한 트리거 알고리즘

이는 기존의 관계형 데이터베이스가 제공하는 도메인의 제약 조건만으로 불가능하므로 트리거를 이용하여 시스템을 구축한다면 해당 테이블 상에서 응용프로그램의 개입 없이 데이터의 무결성을 획득할 수 있고, 서열의 중복성 검출과 버전생성을 자동화 할 수 있다.

5. 결론

국내의 대부분 생물학 연구실에서는 시퀀싱된 서열 파일 정보 관리를 파일형태로 저장해 이용하는 상황이다. 따라서 국내 생물학자들로부터 얻어진 서열 데이터를 자체 관리할 수 있는 서열정보 관리시스템이 요구되었다.

따라서 이 연구에서는 시퀀싱된 유전체 및 단백질 서열 데이터의 효율적인 관리를 위해 유전체와 단백질 서열 정보의 편집, 저장, 검색과 서열 파일 포맷 생성 및 변환을 수행하는 서열 정보관리 시스템을 설계하고 구현하였다. 결과적으로 서열 파일의 변경이나 삭제를 데이터베이스를 이용하여 관리하고 서열 및 버전 서열 데이터의 일관성과 무결성을 유지할 수 있다. 또한 시퀀싱을 통해 생산된 서열 데이터를 트리거를 이용해 효율적으로 저장관리 할수 있고, 국내의 생물 데이터를 장기적으로 축적하고 생물학 및 의학 관련 분야 연구를 촉진할 수 있다.

참고문헌

[1] David W. Mount, "Bioinformatics : Sequence and Genome Analysis" Cold Spring Harbor Laboratory Press, 2001.

[2] G. Stoesser, "The EMBL nucleotide sequence database", Nucl. Acids. Res. 2001.

[3] Y Tateno "DNA Data Bank of Japan in the age of information biology", Nucl. Acids. Res. 1997.

[4] Robin Cover, XML linking and addressing language, Oasis, 2001.

[5] J. Ostell. The NCBI data model. Chapter 2 in Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins, B.F.F. New York: 2001

[6] Dennis A. Benson, "GenBank" Nucl. Acids. Res. 2002