DEVELOPMENT OF XML BASED PERSONALIZED DATAASE MANAGEMENT SYTEM FOR BIOLOGISTS

Kyung Hwan Cho, Kwang Su Jung, Sun Shin Kim, Keun Ho Ryu

Database & Bioinformatics Laboratory, Chungbuk National University, Korea
Cheongju Chungbuk, 361-763, Korea
{khcho,ksjung,sskim04,khryu}@dblab.chungbuk.ac.kr

ABSTRACT:

In most biological laboratory, sequences from sequence machine are stored into file disks as simple files. It will be hard work to store and manage the sequence data with consistency and integrity such as storing redundant files. It is required needed to develop a system which integrated and managed genome data with consistency and integrity for accurate sequence analysis.
There fore, in this paper, we not only store gene and protein sequence data through sequencing but also manage them. We also make a integrate schema for transforming the file formats and design database system using it. As integrated schema is designed as a BSML, it is possible to apply a style language of XSL. From this, we can transfer among heterogeneous sequence formats.

KEY WORDS: Bioinformatics, BSML

1. INTRODUCTION

Through HGP, we are performing the sequencing about the genome and the protein sequence in our country. Single-Primer Sequencing which only once performs the sequencing about Template DNA by the single primer is provided. And also the sequencing by PCR products is served.

Because there is not any software which manages the sequenced sequence information in the laboratory in our country, we just store the data files into the disk and use them. Accordingly, the sequence file can be deleted or modified by the users and also it is not managed consistently. Therefore, it is difficult to manage the integrity and the consistency of the data and to store the sequence data for the long run research.

There are the integrated database retrieval sites such as NCBI[1], EMBL[2], DDBJ[3] and so on in our country. They store the biology data and provide the retrieval service and software. To exchange the sequenced sequence data with other researchers and provide them to the public database, it must be able to change the format freely.

Accordingly, in this paper, we designed and implemented the sequence file information management system. The system edits, stores, retrieves the sequence data and then generates the sequence file format. And it transforms the file, to manage the protein and sequence data efficiently. The sequence management system comprises as follows: Viewer reads from the sequence files and then presents them. And the sequence edit manager processes the sequence operations for editing. And also the sequence data format transformer uses XML which is the standard format for web data, as a common format for exchanging the sequence data. So it can provide the compatibility. And finally, the sequence storing manager designs the database schema to store the informal sequence data into the database.

2. RELATED WORKS

Staden Package[4] is a software for management and analysis of sequenced sequences. It is developed in Medical Research Council Laboratory at Cambridge University. This software consists of four different parts. First, Trev is a viewer of sequenced experiment files. Second, Trace_diff shows the mutation information in the reference data and the trace data. Third, GAP4 shows the sequence assembly and edits the contigue after assembling. Forth, Pregap4 prepares for the data to assemble the sequence data. Fifth, Spin offers a sequence similarity search and operation to analyze the match sequence after assembling. Staden Package supports to analyze and manage the sequence data in the sequence files but doesn't provide the analyzed sequence data to store and retrieve into the database. The sequence file based analysis program can't manage a large amount of data.

NCBI includes many different databases (GenBank as a DNA sequence database, dbEST as a EST sequence database, MMDB as a protein molecular structure modelling database, OMIN as a human gene catalogue and database of genetic variation, and PUBMED as a literature cited database). These heterogeneous databases are linked by hyper links. From these databases, Entrez is used as a searching tool for sequence and gene data. And PubMed is used as a literature searching system and also BLAST is used for sequence similarity searching system.

In GenBank, a lot of identifiers such as Accession Number, Locus, GI(GenInfo Identifier) are used to identify the sequence records. Accession Number is more stable than Locus name. But when we use

Accession to search, it is impossible to search the version sequence. Accordingly, GI which can only identify the sequence in NCBI database is used to identify the sequence data. The sequence which has the same Accession Number is assigned a new GI, whenever the sequence is changed. Therefore, while searching the sequence, it is possible to search all the version sequence and the original sequence. These identifiers are described in ASN.1[5] which is the common format of NCBI.

3. PROPOSED SYSTEM ACHITECTURE

In figure 1, the sequence management system is composed of the sequence file format transformer, the sequence edit manager, and the sequence storing manager, and so on. The sequence format transformer loads the experiment files from the disk in memory. Then EXFileTOXML module transforms them to a common XML format and then keeps the mapping information between XML file and the final files to transform the file format which the users want to. It is composed of the final format generation module which stores and presents the mapping information by XSL.

The experiment file transformed into XML format passes through the sequence editing module and then presents the sequences in the view module. And the sequence operation processor performs a lot of operations about the sequence. The sequence annotator adds the annotation information to store the sequence information.

The sequence storing manager is composed of the sequence version manager, the sequence storing manager, and the sequence information searcher. The sequence version management module checks the versions, when the sequence and the annotation information are inputted into the database. And then it generates a new version and manages the version deletion. The sequence storage manager inputs new sequence version information. And the sequence information searcher receives the retrieved words and then presents the sequence information.

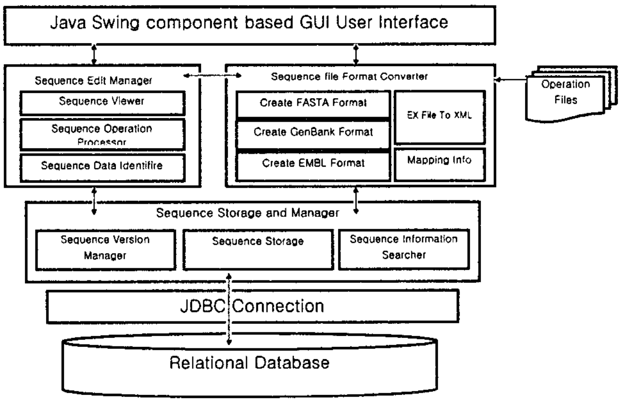

Figure 1. System Architecture

4. EDITION AND TRANSFORMATION OF SEQUENCE DATA

4.1 Sequence editing management

The sequence viewer retrieves the sequences and then presents the base of the selected sequence and the sequence data. The sequence annotator adds the annotation information to the sequence data. And then it stores the protein identifier, the project identifier, the sequence type, the source, the name of the researcher and so on into the database. The sequence operation processor is composed of as follows: Base Composite calculates the composition ratio of the sequence. Set Range selects a special part of the sequence and composes a new entry. Complement Sequence generates the complement sequence of the DNA sequence. Rotate determines the starting position of the sequence. Interconvert operations transform Thymine of DNA to Urasil of RNA while DNA transfers to RNA.

4.2 Sequence file format transformation

Transforming between formats of the heterogeneous information by XML is composed of three modules. First, ExFileToXML extracts the sequence and the related data from the experiment files which include the consistent sequences and the related information through assembling and then describes them into the XML documents. Second, MappingInfo makes the mapping information between the heterogeneous formats and then describes it by XSL. And then it stores in into the database. Third, the XSL mapping information is applied to and then transformed into the format that supports the application of the life information. This system focuses on the sequence data extracted from the laboratory and we defined the experiment files in table 1.

Table 1. definition of Experiment File Format

Sequence ID	: Sequence Identifier
Name	: Name of Nucleic Acid or Protein
Version	: Sequence Version Information
Molecule type	: Sequence of Nucleic Acid or Protein
Sequence length	: Sequence Length
Date	: Create Sequence Date
Source	: Source Organism
DBref	: Identifier of Referenced Database
Base count	: Number of a, c, g, t (Skip when a protein)
Sequence	: Sequence
Experimenter	: Sequencing Experimenter

The annotation line of the FASTA format which we defined represents id of XML document, version, molecule name, molecule type, organism, length field values and so on. And the next line is the sequence data. Figure 2 is the example of FASTA format

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

Figure 2. Example of FASTA Format

Using the experiment file format in table 1, we generate the XML file in figure 3. Seg-set element in XML document is the root element which includes the sequence set. The attributes of the sequence element include the identifiers, the molecule type, the sequence length, the molecule name, the generated date and so on. The attribute elements are added continuously. Currently, they include the source life and the number of the version sequences. The version element represents the version information. And Feature-tables present the element reference information and the feature attributes.

```
<?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   >gtagaaagcaccgacaatactcctggcatgggcgttaaagctcacaggat</seqdata>
 <seqdata type=   complement   >atcctgtgagctttaacgcccatgccaggagtattgtcggtgcttctac</seqdata>
 <version id="AB003468.1     versionNum=1 >
  <vseqdata>ggctgtgtgcacgaacccccgttcagcccgaccgctgcg</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.,
       Masiarz,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>
```

Figure 3. Example of XML File

We extracted the sequence data and the related information from GenBank data file, and then generated the XML document. And then we described the mapping information among GenBank, FASTA, EMBL formats, by XSL. Finally we transformed it into FASTA format, applying XMLTOFASTA.xml. Also we wrote each format by XSL using the mapping information, and then the result can be shown on the web browser.

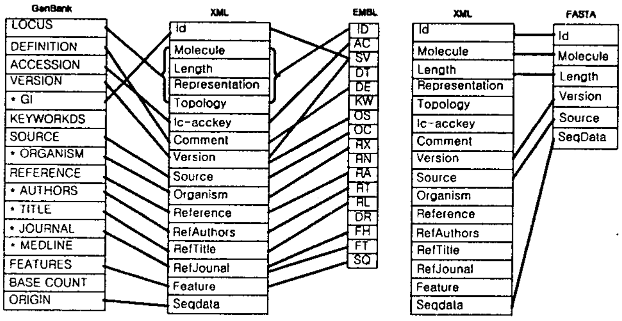

Figure 4. Mapping Information between DNA Sequences Formats

5. STORING AND VERSIONING OF SEQUENCE DATA

5.1 Sequence storing manager

The sequence storing management database stores the protein sequence data and the related information, and performs the experiment of the sequencing and assembly. The sequence data have original identifiers which determine the sequence. There are tables that store the sequence data and the related information. They are Annotation, DerivedSeq, VersionSeq, PSequence, and DSequence. The relational schema of the sequence database is shown in figure 5.

5.2 Sequence version manager applied Trigger

When a new sequence is inputted into the database, the sequence version manager checks if there are the same identifiers in the sequence table, using the trigger. The trigger algorithm is represented as follow.

After checking the sequence version, a new sequence version is inputted into the table, using the trigger in Figure 6.

Project	Pid, Pname, Pdate, Puser	Sequencing experiment Information
Annotation	ASid, Apid, Aog, Aname, Amachine, Adc	Sub Sequence Identifier Information
DSequence	Sid, SPid, Slength, SnumVerSeq, Sdate, Sseq, SnumA, SnumT, SnumC, SnumG	A basic Sequence
PSequence	Sid, SPid, Slength, SnumVerSeq, Sdate, Sseq	Protein Sequence
DerivedSeq	DSid, DPid, Dop, Dseq, Dlength, Dtype, Ddate	Induction Sequence
VersionSeq	VSid, Vid, Vseq, Vlength, Vdate, Vdiffer	Version Sequence

Figure 5. Related Schema for Sequence Database

```
INPUT : Sseq(New Sequence), SID(Sequence ID)
1) Search in Database use a Sequence ID
2) SELECT COUNT(*) INTO Duple FROM Sequence
  WHERE Sid=newrow.Sid AND Sseq=newrow.Sseq;
- IF Duple=0 THEN Check a Overlapping Sequence
- IF Seqtrue=0 THEN Store in Sequence Table
   ELSE COUNT(*) INTO Vertrue FROM VersionSeq
   IF Firstver=0 THEN INSERT VersionSeq
- IF Verture≠0 THEN RAISE_APPLICATION_ERROR
- IF Verture=0 THEN Store Version Table
```

Figure 6. Trigger Algorithm for Version Manage

This system is impossible to construct using the constraint condition of the domain which the existing relational database provides. Therefore we used the trigger to construct the system. We can get the integrity of the data on the table without the application program. And also we can check the redundancy of the sequence and automate to generate the version.

6. CONCLUSIONS

Most of the biology laboratory in our country manages and stores the sequenced sequence file in the form of file. Accordingly, it is required to have the sequence information management system which can manage the sequence data collected by the biologist in our country.

Accordingly, in this research, we designed and implemented the sequence file information management system. The system edits, stores, retrieves the sequence data and then generate the sequence file format and transform it, to manage the protein and sequence data efficiently. Consequently, it is possible to modify and delete the sequence file in the database. And also it can be able to keep the integrity and the consistency of the sequence and version sequence data. And we can store and manage efficiently the sequence data generated through sequencing by the trigger. We can collect the long term biology data and develop the biology and medical science accordingly.

7. ACKNOWLEDGEMENTS

This work was supported by the Regional Research Centers Program of Ministry of Education & Human Resources Development in Korea.

REFERENCES

[1] R.Srikant, R.Agrawal, 1995, "Mining Generalized Association Rules", VLDB

[2] Steve Pepper, 2000, "The TAO of Topic Maps", XML 2000 Conference & Exposition

[3] A.Maedche, S.Staab, 2000, "Semi-Automatic Engineering of Ontologies from Text", Institute AIFB, Karlsruhe University, Germany

[4] S.Pepper, G.Moore, 2001, "XML Topic Maps(XTM) 1.0", TopicMaps.Org

[5] A.Maedche, S.Staab, 2001, "Discovering Conceptual Relations from Text", Institute AIFB, Karlsruhe University, Germany

[6] D.Braga, A.Campi, S.Ceri, M.Klemettinen, P.Lanzi, 2003, "Discovering interesting information in XML data with association rules", SAC

[7] Q.Ding, K.Ricords, J.Lumpkin, 2003, "Deriving General Association Rules from XML Data", SNPD

[8] Jacky W.W.Wan, G.Dobbie, "Mining Association Rules from XML Data using XQuery", 2004, ACM International Conference Proceeding

[9] Hyo Soung Cha, Kwang Su Jung, Young Jin Jung, Keun Ho Ryu, 2004 "Building an Integrated Protein Data Management System Using the XPath Query Process", of the 20th ISRS conference

[10] Hyo Sung Cha, Kwang Su Jung, Young Jin Jung, Keun Ho Ryu, 2004 "Building a Biological Genomic Database Management System in Laboratory Level", of the 31th KISS spring conference, Vol.21 No.2, pp52-54

[11] Sung Hee Park, Kwang Su Jung, Young Jin Jung, Hyo Sung Cha, Young Uk Kim, Keun Ho Ryu, 2004 "A Personalized Biological Database System based on XML", ISMB/ECCB 2004 Poster, Scottish Exhibition & Conference Center, Glasgow, Scotland, UK

[12] Young Uk Kim, Kwang Su Jung, Young Jin Jung, Hyo Sung Cha, 2004 "The Biological Data Converter based on BSML for Sharing Information", of the 31th KISS fall conference