Building an Integrated Protein Data Management System Using the XPath Query Process

Hyo Soung Cha[0], Kwang Su Jung, Young Jin Jung, Keun Ho Ryu
Database Laboratory, Chungbuk National University
Cheongju, Korea
{kkido, ksjung, yjjeong, khryu}@dblab.cbu.ac.kr

Abstract: Recently according to developing of bioinformatics techniques, there are a lot of researches about large amount of biological data. And a variety of files and databases are being used to manage these data efficiently. However, because of the deficiency of standardization there are a lot of problems to manage the data and transform one into the other among heterogeneous formats. We are interested in integrating, saving, and managing gene and protein sequence data generated through sequencing. Accordingly, in this paper the goal of our research is to implement the system to manage sequence data and transform a sequence file format into other format. To satisfy these requirements, we adopt BSML (Bioinformatics Sequence Markup Language) as the standard to manage the bioinformatics data. And then we integrate and store the heterogeneous flat file formats using BSML schema based DTD. And we developed the system to apply the characteristics of object-oriented database and to process XPath query, one of the efficient structural query, that saves and manages XML documents easily.

Keywords: Bioinformatics, BSML, XPath query.

1. Introduction

At the present day, Bioinformatics is developing rapidly. It is the application to automate, computerize, and analyze the biology data using a new technology in computer science and statistics.

With bioinformatics development, it is necessary to manage and operate a large amount of data that have been collected till now. Accordingly, the standard related to biology[1] is organized to make a variety of biology data exchange easily.

Nowadays sequencing service is provided on the web. But there isn't any software to manage the sequenced sequence file data in most of the domestic laboratory of biology. So they have been stored in the form of file. As a result sequence data are not managed consistently. Therefore, it is difficult to manage biology data with integrity and consistence.

In this paper, recently because the database for only XML is used widespread, we can construct the database for gene data. And we developed the software to transform, edit, store, and retrieve the biological data between heterogeneous flat files based BSML, XML format, for gene data. These sequences can make us share information between biology data more efficiently, manage and edit the sequence of version through experiments. And the data are stored and retrieved in the form of BSML.

This system is able to share the information between biology data more efficiently. And we used the database for only XML documents to reduce time and cost consumption. Because the existing relational database stores tree based XML documents into the flat table or integrated schema table, inconsistency would be occurred. When the documents are retrieve, join operation with high cost is needed as well. However, the object database designed, based on XML structure attributes is efficient, when tree based XML documents are stored. Also, we can reduce the cost and effort to parse and store the XML documents in the relational database. Therefore we need XPath query, not relational query, which we can find out the information that we wanted to get. And it is also efficient, when we manage the ordered and structural document information.

2. Related Works

There have been many researches to integrate and manage the information, because the data of life information are stored just in the form of flat file. Data integration techniques are classified into the following three ways. First, it is a link based integration technique used most widespread recently. It links the flat file database using WWW links and index. But because it is not a real linkage, the potential of error is high and it could not support ad-hoc query. Second, it is a meta-data based technique. It generates the integrated view about data source and processes the query by the integrated query mechanism. But it has some disadvantages, that is, the response time through Internet takes so long and the data reorganization is so difficult. Third, it is the data warehouse based technique. It generates the integrated schema, loads, and manages all the source data into the data warehouse, according to the generated schema. The waiting time is not needed and the dependence on the Internet is so low that the confidence about system is very high.

The proposed BSML[2] to exchange the biology data between biology information systems encodes DNA information differently from AGAVE, BIOML, DAS, and GAME. And it is more specified in the expression way than other format in XML. The BSML DTD started in 1997 was updated into version 3.1 in 2002. A lot of application program databases are using it to exchange gene data and visualized protein data. The BSML has the advantages represented above. And it was adopted a new standard to represent the bio sequence data in American National Biology Information Center such as NCBI[3], EMBL[4], PDB[5], and so on. Therefore, in this research we constructed the integrated schema using BSML DTD on the data warehouse integrated by technique of BSML.

3. System Modeling

The architecture of the system is shown in Figure 1. First, there is Oracle 9.2.20.4.0 XML DB containing source data related protein data with BSML DTD in the bottom of the system. Second, there is Integrated Data Format Transformer, which transforms many kinds of flat file into an integrated schema and then stores into object class or file. Third, there is MyPage Editor, which edits or stores annotations or versions of sequences. Fourth, Biology Information Retrieval Query Processor parses the source data and processes the query from XPath query. Finally, Storage Manager stores and manages the data.

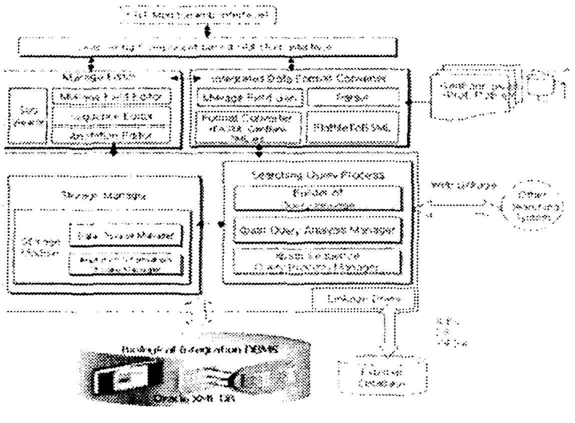


Fig. 1 The structure of integrated protein data management system using the ORDBMS

4. Integrated Protein Management System

1) Integrated Data Format Transformer

It is possible to parse flat files and then generate BSML schema in Figure 3, where the flat files are GenBank, PDB, Swiss-Prot, and FASTA provided from the representative database of life information. Format Transformer is composed of several modules such as parsing module, FlatFileToBSML module, format transformer, and MyPage field generator. First, data are extracted in the parsing module. FlatFileToBSML module stores the data in the form of object and then describes them into XML documents depending on BSML DTD and integrated schema structure. The Format Transformer makes objects or XML data to recover to flat files such as GenBank or FASTA according to generated BSML structure.

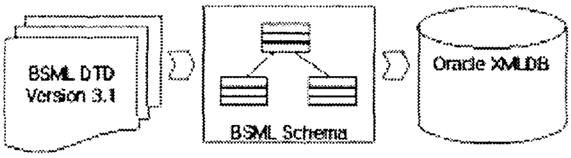
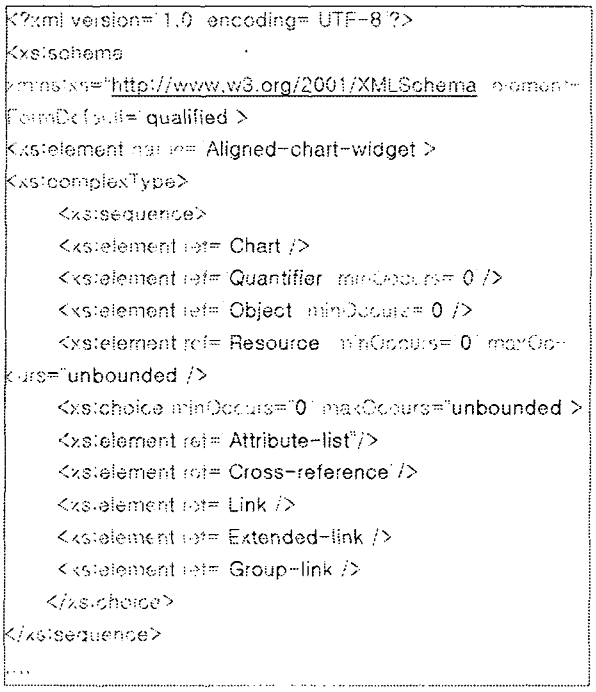


Fig. 2 Transformed BSML schema and relation of XMLDB

Figure 2 shows transformation steps to register the schema of BSML DTD in the relational database.



```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema
xmlns:xs="http://www.w3.org/2001/XMLSchema" elementF-
ormDefault="qualified">
<xs:element name="Aligned-chart-widget">
<xs:complexType>
     <xs:sequence>
     <xs:element ref="Chart"/>
     <xs:element ref="Quantifier" minOccurs="0"/>
     <xs:element ref="Object" minOccurs="0"/>
     <xs:element ref="Resource" minOccurs="0" maxOc-
curs="unbounded"/>
     <xs:choice minOccurs="0" maxOccurs="unbounded">
     <xs:element ref="Attribute-list"/>
     <xs:element ref="Cross-reference"/>
     <xs:element ref="Link"/>
     <xs:element ref="Extended-link"/>
     <xs:element ref="Group-link"/>
    </xs:choice>
</xs:sequence>
...
```

Fig. 3 Transformed BSML schema based on DTD

```
BEGIN
  DBMS_XMLSCHEMA.registerSchema(
   'http://www.w3.org/2001/XMLSchema/bsml3_1.xsd',
    getclobdocument('bsml3_1.xsd'),  TRUE.  TRUE.  FALSE.
FALSE
  ):
END;
CREATE TABLE BSMLTABLE(
NAME VARCHAR2(30) PRIMARY KEY,
DOC XMLTYPE
) XMLTYPE COLUMN DOC
  XMLSCHEMA
  "http://www.w3.org/2001/XMLSchema/bsml3_1.xsd"
  ELEMENT "Aligned-chart-widget"
insert into BSMLTABLE values('bsml00_vdata.xml',
XMLType(getClobDocument('bsml00_vdata.xml')))
```

Fig. 4 The process that is registering the schema, creating the table, and storing the documents into the XML DB

Figure 4 could present the SQL query processing to store in the generated table, after the transformed Oracle XML DB is registered and the table is generated.

2) MyPage Editor

Sequence viewer module selects the retrieved information generated through XPath query and shows the quantity of selected sequence A, C, G, T and sequence information or annotation information. MyPage file editor could be created through integrated data format transformation and then parses or process the query. Then MyPage field module will be activated. Sequence editor[6] is composed of five operations. First, Base Composite operation computes the ratio of base sequence. Second, Set Range operation generates a new entry appointing both the starting point and end point of a particular part. Third, Complement Sequence operation generates complement sequence. Fourth, Rotate operation operates rotation. Last, the process of transcription between DNA and RNA, that is, rotate operation converts Thymine(T) and Uracil(U) mutually. Annotation editor adds the annotation information necessary for editing the data and stores it.

Figure 5 shows the process of BSML document generation and storage in the relations of modules between MyPage editor and Storage Manager.

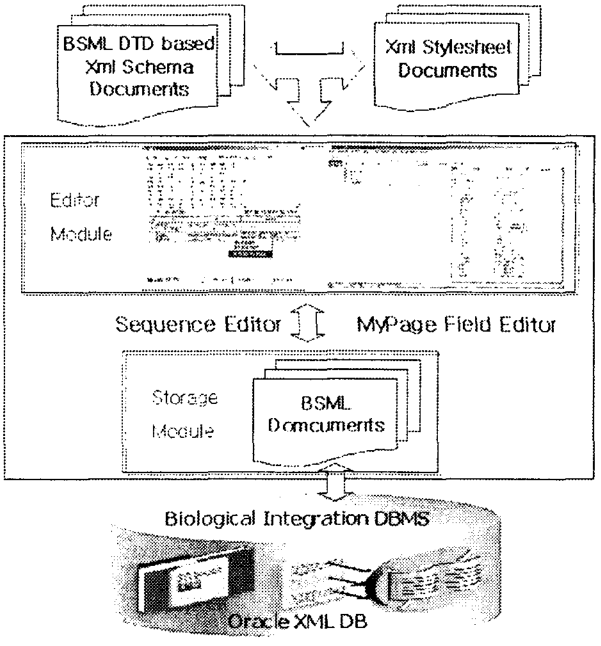


Fig. 5 The relations of modules between MyPage editor and Storage Manager

3) Biological Information Retrieval Query Processor and Storage Manager

Biology Information Retrieval Query Processor is consist of Retrieval Query Language Processor, XPath Query Parser, XPath Sequence Query Processor, Retrieval Result Browsing, and so on. User inputs the query through Retrieval Query Language Builder. Then it invoke a query to XML DB using XPath query and then the result will be browsed. Also, the storage manager consists of Biology data storage and annotation information storage.

```
select extract(e.doc,'chs:GI//feature',
'xmlns:chahyosung="http://www.w3.org/2001/XMLSchema/"'),
extract(e.doc,'chahyosung:Aligned-chart-widget//humen'.
'xmlns:chahyosung="http://www.w3.org/2001/XMLSchema/"')
from BSMLTABLE e
where
existsNode(e.doc,'chs:Aligned-chart-
widget//feature[title="source"]'.
'xmlns:chahyosung="http://www.w3.org/2001/XMLSchema/"')=1
and
existsNode(e.doc,'chs:Aligned-chart-
widget//feature[title="mise_feature"]'.
'xmlns:chahyosung="http://www.w3.org/2001/XMLSchema/"')=1
```

Fig. 6 The XPath Query about the stored BSML

Figure 6 shows the query about stored BSML documents. Because XPath query has a complicated structure, we implemented the interface convenient for the user, using schema reference query.

5. Implementation

There were lots of errors in XML DB when we used Oracle 9.2.0.1.0 version. So we constructed database after patching Oracle 9.2.0.4.0 version. We reference to Java Document Model(DOM) API in Oracle company to store XML Type.

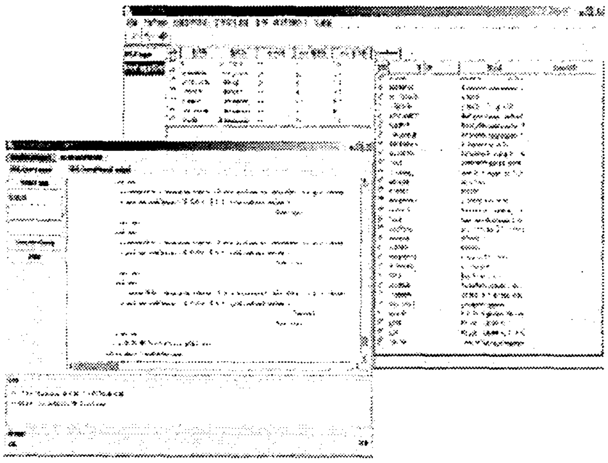


Fig. 7 The system interface and retrieval interface using XPath

Figure 7 shows the system interface and retrieval interface using XPath. At the back, there is MyPage editor on the left. And on the right, we can see the parsed flat files matched and viewed in BSML schema structure. Accordingly, necessary objects are moved into MyPage editor on the left and then the information can be managed and retrieved using XPath query.

6. Conclusion

Though a lot of efforts have been tried to standardize the biology information files, a large amount of bio data have not been managed and stored in bioinformatics. Also there is an essential limitation to design RDBMS to represent XML data with hierarchical structure in the relational database, which is composed of two-dimensional tables. So in this research, to manage the integrated protein data, we used the object relational database and stored them in BSML schema based on BSML DTD. And then we implemented the retrieval system utilizing XPath query process. MyPage editor parses a variety of flat file formats and then edits and modifies them. And after the editor generates a new BSML, it stores and manages BSML. And it's possible to retrieve the data efficiently through range or structural query, as using XPath query process. As a result, it's possible to accelerate to develop the system related structure such as protein structure, bio Pathway, bio ontology, and so on.

References

[1] URL : http://www.visualgenomics.ca/gordonp/xml/

[2] URL : http://www.bsml.org/

[3] URL : http://ncbi.nlm.nih.gov/

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

[5] URL : http://www.rcsb.org/pdb/

[6] P. Sung-Hee, R. Keun-Ho, 2002, Building Genome and Protein sequence information Management System. KOSTI.

[7] URL : http://otn.oacle.com/software/index.html

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

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

[10] Frederic Achard, Guy Vaysseix, Emmanuel Barillot, 2001, XML-Bioinformatics and data integration, ISMB, Vol.17 no.2 p115-125

[11] Bonfield, James, K., Beal, Kathryn F., Betts, Matthew J. and Staden, 2002, Trev: a DNA trace editor and viewer. Bioinformatics Vol.18, pp194-195.

[12] J. Kwang Su, P. Sung-Hee, R. Keun Ho, Hyeon S. Son, 2002, Sequence Version Management System based on Trigger, Korean Society for Bioinformatics Annual Meeting, Vol.1, pp134-141.

[13] L. Rong Hua, P. Sung-Hee, J. Byeong-Jin, R. Keun Ho, 2002, Transformation of heterogeneous data files for bioinformatics. Korean Society for Bioinformatics Annual Meeting, Vol.1, pp118-124.

[14] Shurug Al-Khalifa, Cong Yu, H.V.Jagadish, 2003, Querying Structured Text in an XML Database, SIGMOD2003, San Diego.