

Mapping Biological XML DTDs Using Ontology

Rana Hashmy

Systems Analyst, Post Graduate Dept. of Computer Science
University of Kashmir, Srinagar – 190006 (J&K), India
ranahashmy@gmail.com

Abstract

Several biological databases exist which use different formats for storing data. Further, each database has its own schema and a query interface. There exist no standard conversion tools for converting data from one format to another, which makes querying multiple heterogeneous databases, a difficult task. Since XML provides a standard format for sharing and exchanging data on WWW, it is used to share information in biological databases. Various DTDs are defined for biological databases but they do not have a unified format. Same data is being represented as different formats using different DTDs. This paper provides mapping between various biological XML DTDs and ontology so that it is possible to correlate the data between the DTDs.

Keywords: Ontology, XML, biological databases

1. Introduction

Traditionally, biological data is in a flat file format. With XML being adapted as a powerful interchange language for data among different applications, attempts have been made to describe biological data using XML. The goal of developing XML DTD [11] for biological databases has been to provide an extensible framework and to facilitate exchange of information between scientists all over the world through the World Wide Web. There are various DTDs for biological databases, like, BIOML [6], ProML [1], CML [9], GAME [10], AGAVE [3], and BSML [8]. The problem with these DTDs is that they don't have a uniform format. Mostly, similar data is being represented as different formats in different DTDs. There is a need to develop such a system that would facilitate scientists/biologists to be able to query biological databases more efficiently, and would

provide a unified platform to access various kinds of biological data. To address this kind of issue, ontology is becoming popular because it provides a common vocabulary in which biological data can be expressed [7].

We are providing a mapping between various DTDs using the ontological concepts as the common terms for them, hence developing a relation between various DTDs. We are using BAO [5, 2], which is a three-dimensional domain ontology for Biological and Chemical Information Integration system BACIIS [4]. This system integrates Life Sciences databases mainly Swissprot [12], GenBank [13], PDB [14], OMIM [15].

Attempts have been made to distribute biological data in various XML formats. XEMBL [18] is a tool for distributing EMBL [16] data in XML format. Currently it distributes data in two formats, AGAVE and BSML. It takes EMBL accession number as input to query data, and provides with an option to select the output format. The limitation here is that it may not always be possible for the user to know and remember all the accession numbers related to the proteins and genomes. Also, the complete information for a given accession number for a protein or a genome is displayed. It does not facilitate querying data with conditions as predicates. Our work differs from this in that we can use a more generic approach by replacing each tag of the XML formats by the ontological terms in the two DTDs, hence providing a mapping or a relation between these two DTDs. Later, these generic ontological terms, which are more meaningful to the end user, can be used to query data.

2. Analysis of the Problem

As has been brought out above, the DTDs have been developed with varied purposes. As a result more than one DTD can be used to describe data in a single biological database. For example, the DTDs, BSML

and AGAVE, can be used to describe data in SwissProt. Since the DTDs are different, the representation of the same information is different in each DTD. As an example a part of the AGAVE and BSML DTDs are shown in Figure 1.

```

<! ELEMENT db_id EMPTY>
<! ATTLIST db_id
    id CDATA #REQUIRED
    version CDATA #IMPLIED
    db_code CDATA #REQUIRED >
Example from AGAVE DTD

<! ELEMENT Attribute EMPTY >
<! ATTLIST Attribute
    name CDATA #REQUIRED
    content CDATA #IMPLIED>
Example from BSML DTD

```

Figure 1

In AGAVE *db_id* is an identifier for an object in its source database, *id* is a data identifier such as GenBank accession or PID, *db_code* is a code for the data source, e.g. GenBank is “gb”, *version* is the version of the associated data.

BSML represents the same information expressed in AGAVE through the element named as *Attribute*. *name* contains the list of cross-references, and *content* defines data identifier and data source.

A portion of the data according to AGAVE DTD with reference to Figure 1 is expressed as follows:

```

<db_id db_code = “SWISSPROT”
    id = “O26117”/>

```

Same data according to BSML DTD is expressed in the following manner:

```

<Attribute name = “database-xref”
    content = “SWISSPROT:O26117” />

```

It can be seen from this example that a variable in BSML is expressed here as a combination of two variables in AGAVE.

Another example of a different kind of dissimilarity between the two DTDs expressing same data can be seen in figure 2.

According to Figure 2, data in AGAVE format is expressed as:

```

<keyword>
MEDLINE:      98037514:      PUBMED:
9371463:database: MEDLINE, PUBMED:
authors: Smith D.R., Doucette-Stamm L.A.,
Deloughery C., Lee H., Dubois J., Aldredge T.,
Bashirzadeh R., Blakely D., Cook R., Gilbert
K., Harrison D., Hoang L., Keagle P., Lumm
W., Pothier B., Qiu D., Spadafora R., Vicaire
R., Wang Y., Wierzbowski J., Gibson R.,

```

```

Jiwani N., Caruso A., Bush D., Reeve .N.: title:
Complete genome sequence of
Methanobacterium thermoautotrophicum
deltaH: functional analysis and comparative
genomics: type: Journal of Bacteriology
179(22):7135-7155(1997):

```

```

</keyword>

```

```

<!ELEMENT bio_sequence (db_id , note? , description?,
keyword*, sequence? , alt_ids? , xrefs? ,
sequence_map*, map_location*)>
<! ELEMENT keyword (#PCDATA)>
Example from AGAVE DTD

<! ELEMENT Reference (Attribute*, RefAuthors?,
RefTitle?, RefJournal?)>
<! ATTLIST Reference %attrs;
    dbxref CDATA #IMPLIED
    refs IDREFS #IMPLIED >
<! ELEMENT RefAuthors (#PCDATA)>
<! ELEMENT RefTitle (#PCDATA)>
<! ELEMENT RefJournal (#PCDATA)>
Example from BSML DTD

```

Figure 2

BSML format expresses the same data as follows:

```

<Reference dbxref="98037514">
  <Attribute name="cross-reference"
    content="MEDLINE; 98037514" />
  <Attribute name="cross-reference"
    content="PUBMED; 9371463" />
  <RefAuthors>Smith D.R., Doucette-Stamm L.A.,
  Deloughery C., Lee H., Dubois J., Aldredge T.,
  Bashirzadeh R., Blakely D., Cook R., Gilbert K.,
  Harrison D., Hoang L., Keagle P., Lumm W.,
  Pothier B., Qiu D., Spadafora R., Vicaire R.,
  Wang Y., Wierzbowski J., Gibson R., Jiwani N.,
  Caruso A., Bush D., Reeve J.N.
</RefAuthors>
  <RefTitle>Complete genome sequence of
  Methanobacterium thermoautotrophicum deltaH:
  functional analysis and comparative genomics
</RefTitle>
  <RefJournal>Journal of Bacteriology 179(22):
  7135-7155(1997)
</RefJournal>
</Reference>

```

In this example one single element *keyword* in AGAVE is split as multiple elements *Attribute*, *RefAuthors*, *RefTitle*, *RefJournal* in BSML.

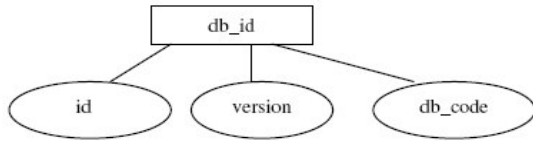


Figure 3(a)

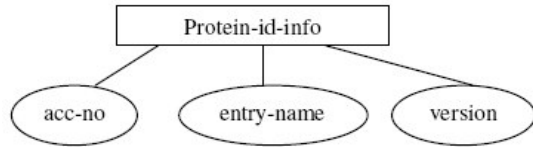


Figure 3(b)

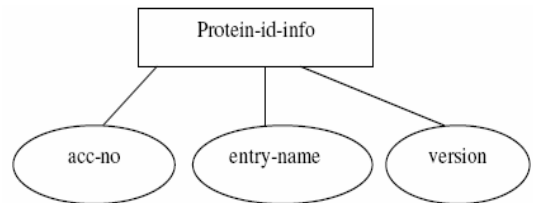


Figure 3(c)

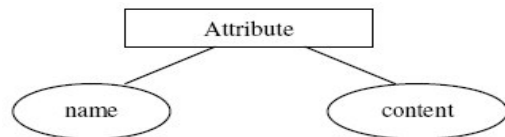


Figure 4(a)

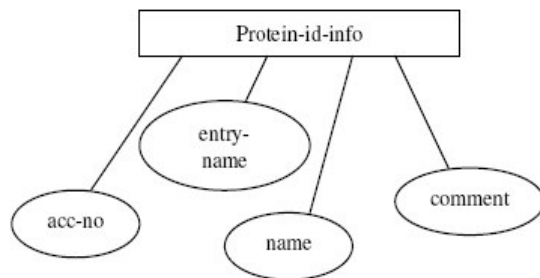


Figure 4(b)

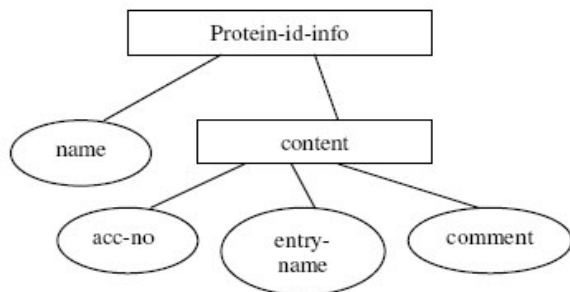


Figure 4(c)

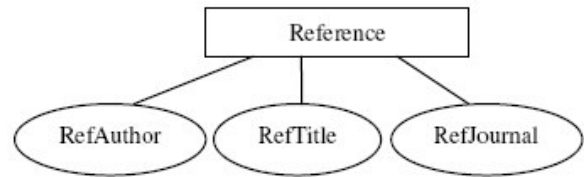


Figure 5(a)

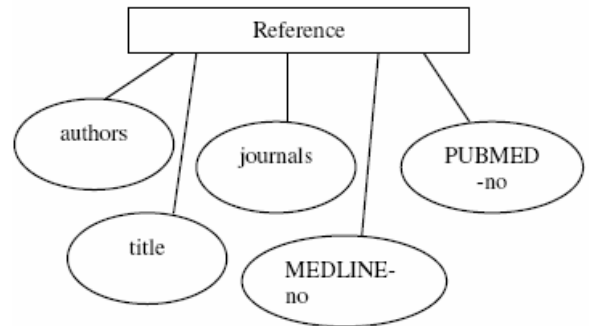


Figure 5(b)

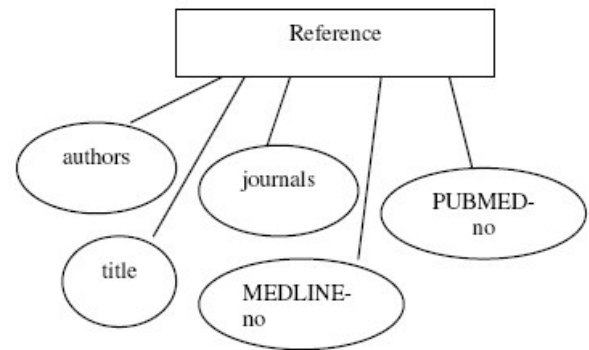


Figure 5(c)

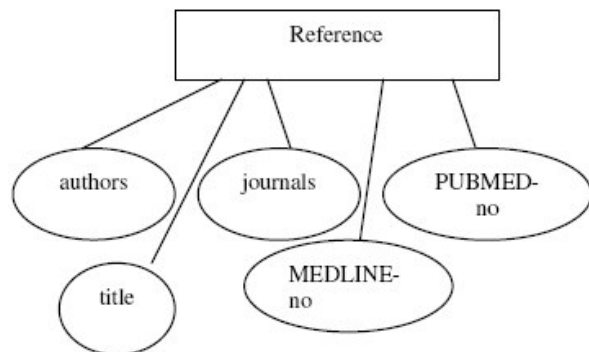


Figure 6

4. Solution

This paper provides a mechanism for identifying similar information in more than one DTD using BAO, as a solution to the problem discussed above.

In figure 1, for AGAVE, *db_id* is an element with attributes *id*, *version*, *db_code*. This has been represented in pictorial form in figure 3(a).

In BAO, if we look at the substructure of protein [5], the ontological concept *protein-id-info* has subclasses of properties as *acc-no*, *entry-name*, *name*, *comment*. This is shown in pictorial form in figure 3(b). This represents the same information as that expressed in above example of AGAVE. Now the mapping between BAO and AGAVE can be done as follows. We replace *id* with *acc-no*, *db_code* with *entry-name* the information contained in *id* and arrive at figure 3(c) with the required mapping.

Now, in figure 1, from the example of BSML DTD we see that *Attribute* is an element with attributes *name*, *content*. This is represented in pictorial form in figure 4(a). This represents the same information as that expressed in figure 4(b) for BAO. Now the mapping between BAO and BSML can be done as follows. We change *content* into element, with *acc-no*, *entry-name* and *comment* as attributes of *content*. Thus we arrive at figure 4(c) with the required mapping.

In Figure 2, for BSML, *Reference* is an element with child elements *RefAuthor*, *RefTitle*, *RefJournal*. This has been represented in pictorial form in figure 5(a).

In BAO, in the substructure of protein, the ontological concept *Reference* has subclasses of properties as *authors*, *journals*, *title*, *MEDLINE-no*, *PUBMED-no*. This is shown in pictorial form in figure 5(b). This represents the same information as that expressed in above example of BSML. The mapping between BAO and BSML can be done as follows. We replace *RefAuthor* with *authors*, *RefTitle* with *title*, *RefJournal* with *journals*, and arrive at figure 5(c) with the required mapping.

Now, in figure 2, from the example of AGAVE DTD we see that *keyword* is an element. This represents the same information as that expressed in figure 5(b) for BAO. The mapping between BAO and AGAVE can be done by replacing *keyword* with *Reference* as shown in figure 6.

More cases can be identified to bring out other dissimilarities between the two DTDs.

From the given examples it is clear that one can establish mapping between different DTDs for correlating the data through the use of BAO ontology.

5. Conclusion

In this paper we have established mapping between the two DTDs – AGAVE and BSML, for correlating the data through the use of BAO ontology.

Hence, given an ontological term for one XML data we are able to provide information from more than one DTD in their given format. Later, we will also be able to design a dynamic query interface so that the user will not have to remember various formats of DTDs. It would be possible to write a generic query to retrieve data in different formats.

References

- [1] Hanisch D., Zimmer R., and Lengauer T., ProML—the Protein Markup Language for specification of protein sequence, structures and families.
- [2] Zina Ben-Miled, Yue W. Webster, Yang Liu, Nianhua Li: An Ontology for Semantic Integration of Life Science Web Databases. *Int. J. Cooperative Inf. Syst.* 12(2): 275-294 (2003).
- [3] AGAVE (Architecture for Genomic Annotation, Visualization and Exchange), <http://www.agavexml.org/>
- [4] Z. B. Miled, O. Bukhres, Y. Wang, N. Li, M. Baumgartner and B. Sipes, Biological and chemical information integration system, Network Tools and Applications in Biology Genoa, Italy, May 2001.
- [5] Z. Ben-Miled, Y. W. Webster, N. Li, O. Bukres, A. K. Nayar, J. Martin and R. Oppelt, BAO, A biological and chemical ontology for information integration, *Online J. Bioinformatics 1* (2002) 60-73.
- [6] The Biopolymer Markup Language, <http://xml.coverpages.org/bioml.html>
- [7] <http://www.w3.org/TR/webont-req/#onto-def>.
- [8] Bioinformatics Sequence Markup Language, <http://xml.coverpages.org/bsml.html>
- [9] Chemical Markup Language, www.xml-cml.org
- [10] Genome Annotation Markup elements (GAME) - <http://xml.coverpages.org/game.html>.
- [11] XML DTD/Schema - www.w3.org/TR/xmlschema-0/
- [12] A. Bairoch, SWISS-PROT Protein Knowledgebase User Manual, 2001.
- [13] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L. (2006) *GenBank Nucleic Acids Res*, 34, D16–D20
- [14] The RCSB Protein Data Bank - www.rcsb.org/pdb/
- [15] OMIM - Online Mendelian Inheritance in Man - www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
- [16] www.ebi.ac.uk/embl
- [17] Okubo, K., Sugawara, H., Gojobori, T., Tateno, Y. (2006) DDBJ in preparation for overview of research activities behind data submissions *Nucleic Acids Res*, 34, D6–D9
- [18] Lichun Wang, Jean-Jack Riethovan, and Alan Robinson, XEMBL: Distributing EMBL Data in XML format, *Bioinformatics Applications Note Vol. 18 No. 8* (2002) 1147-1148