Chapter 7

Conclusion and Future Work

6.1 Conclusion

In chapter 1 we discussed biological databases, bio-ontologies and biological XML DTDs in detail. Biological databases represent an extraordinarily diverse collection of biological data which is increasing rapidly. The functions of biological database are to make biological data available to scientists, and to make biological data available in computer-readable form.

The main characteristics of biological database are that the data stored is highly diverse in nature and is represented heterogeneously. Also different sources offer different interfaces and querying capabilities. There are various types of biological database according to different characteristics of biological data. We have discussed various databases like UniProtKB/Swiss-Prot, EMBL, GenBank, PDB, OMIM, ACeDB, PROSITE, EC-ENZYME, GDB, PIR-PSD, MGD, MEDLINE, Stanford Micro Array Database (SDB), KEGG, Pfam, FlyBase, and the Structural Classification of Proteins (SCOP) database.

Biological databases are stored in various formats like flat file format, relational database format and object-oriented database format. Problem faced in these formats is that the data cannot be shared and exchanged online amongst the scientists. To overcome this difficulty XML was adapted for sharing and exchanging data over the web. Various XML DTDs have been proposed to represent biological data according to their different properties. We have discussed some popular XML DTDs like PROML, BSML, AGAVE, BIOML, CML, and GAME in detail.
Due to the heterogeneous nature of biological data bioinformatic sources are difficult and time-consuming for biologists to use in combination with one another. Various systems have been proposed to integrate and/or query these heterogeneous databases. There are various querying and integration approaches for these databases like warehouse integration, navigational integration, mediator-wrapper approach, and federated database systems. We surveyed some of the existing systems based on these approaches like Sequence Retrieval System, K2/BioKleisli, Transparent Access to Multiple Bioinformatics Information Sources (TAMBIS), Biological and Chemical Information Integration System, [37], XEMBL, Light-weight Integration of Molecular Biological Databases (LIMBO), and [40] in chapter 2.

The problem of sharing and exchanging data online was overcome with the advent of XML, but the problem of storage of biological data in heterogeneous format continued to persist. To address this issue, ontology became popular because it provides a common vocabulary in which biological data can be expressed. We discussed Gene Ontology (GO) and domain ontology for BACIIS in detail in chapter 1. After studying various XML DTDs we observed that in some cases same biological data has been represented in different formats using different DTDs. Because of this reason there is a need to establish relation between various DTDs. Using the concepts of ontology we have provided mapping of XML DTDs with BAO in chapter 3. Here the mapping has been done manually. In manual mapping, mapping is established by looking at terms in the DTD and finding the corresponding term in BAO. This mapping is stored in a table.

We improvised upon this in chapter 4 by providing an automatic mapping between XML DTDs using RDF documents of BAO. The Resource
Description Framework (RDF) is a language for representing information about resources in the World Wide Web. It represents metadata about Web resources, and uses XML syntax. The main design goals of RDF are to have a simple data model, to have formal semantics, to use an extensible URI-based vocabulary, to use an XML-based syntax, to support use of XML schema datatypes, and to allow anyone to make statements about any resource.

To achieve automatic mapping between the XML DTDs we have used algorithms discussed in paper [68] by Lu Xiao et al. In this paper they have achieved the mapping by defining various terms like ontology graphs, DTD structure trees, full covering of DTD structure trees and half covering of DTD structure trees. They have maintained a synonym table manually with the help of domain expert to search the corresponding elements of ontology in DTD.

In our proposed system, INQUES, we have generated synonym table automatically for the biological data through the RDF documents of BAO. We have applied the algorithms discussed in paper [68] to the tree structure of the XML DTDs, and RDF documents of BAO.

In chapter 5 we have proposed a querying system to query multiple biological databases expressed in XML format. With our proposed architecture we are able to achieve integration of various biological databases through the use of ontology. We have been able to overcome the drawbacks discussed for the existing systems by providing a mapping between various DTDs and ontology, thereby, providing the flexibility of defining our input for querying data source. We are able to pose query in different formats for different databases. The query output can be displayed in terms of ontology, or desired XML format.
In chapter 6 we describe how we are achieving integration of multiple biological databases expressed in XML format through INQUES.

6.2 Future Work

Through INQUES we have achieved the querying and integration of heterogeneous biological database expressed in XML format with the use of domain ontology of BACIIS, BAO. This system can be further extended for querying within context. Also it can be used to query multiple databases for a particular query, and instead of displaying cross references in the result of that query, it should display the extracted information from those cross-references in a systematic manner.

Currently our system integrates and queries various biological databases expressed in XML format. In addition to the above future work we should also be able to extend this system for integration and querying of heterogeneous biological data sources.