A System to Integrate and Manipulate Protein Database Using BioPerl and XML

Zurinahni Zainol, Rosalina Abdul Salam, Rosni Abdullah, Nur'Aini, and Wahidah Husain

Abstract—The size, complexity and number of databases used for protein information have caused bioinformatics to lag behind in adapting to the need to handle this distributed information. Integrating all the information from different databases into one database is a challenging problem. Our main research is to develop a tool which can be used to access and manipulate protein information from difference databases. In our approach, we have integrated difference databases such as Swiss-prot, PDB, Interpro, and EMBL and transformed these databases in flat file format into relational form using XML and Bioperl. As a result, we showed this tool can search different sizes of protein information stored in relational database and the result can be retrieved faster compared to flat file database. A web based user interface is provided to allow user to access or search for protein information in the local database.

Keywords—Protein sequence database, relational database, integrated database.

I. Introduction

OUR major activity in Bioinformatics is the management of the database which contains DNA, RNA or protein data. These databases are used by biologists to manipulate and analyze data. One of the basic operations that need to be implemented on the protein database is to find the protein sequence. Currently, the local scientists use BLAST [15] or FASTA to find the protein sequences and other information in the public domain database.

Public sequence databases contain one of the most frequently accessed information on the web [2]. Databases such as Swiss Prot and EMBL provide descriptions of common biological entities (genes, proteins, among others). Numerous computational methods and algorithms in data mining have been implemented to extract hidden knowledge in these databases.

Manuscript received May 19, 2005. This work was supported by University Science Malaysia (USM) under USM short term grants no 304/PKOMP/ 635070.

Zurinahni Zainol is Senior Lecturer at the School Computer Science, USM, Malaysia (e-mail: zuri @cs.usm.my).

Rosalina Abdul Salam is Senior Lecturer at the School Computer Science, USM, Malaysia (e-maill: rosalina @cs.usm.my).

Rosni Abdullah is Associate Professor at the School Computer Science,USM, Malaysia (e-mail: rosni@cs.usm.my).

Nur'Aini is Senior Lecturer at the School Computer Science, USM, Malaysia (e-mail: nur'aini@cs.usm.my).

Wahidah Husin is Senior Lecturer at the School Computer Science, USM, Malaysia (e-mail: wahidah@cs.usm.my).

Therefore, the problem of managing enormous databases is compounded by the fact that if the scientist wants to receive the maximum benefit from this information, they must compare and relate information from difference databases. A new standard is needed to interrogate biological data stored in different types of data presentation [1].

Further more; standardized data formats(s) are also needed to exchange data because all the information is stored in different format databases. Our approach will integrate all different protein databases into once central database for easier and faster access by end users. Besides that our system will also provide a tool enabling data management and search facilities.

II. Related Work

In order to integrate the different protein databases, data exchange standard is needed to integrate these databases. There are many standards currently implemented such as flat files, common Object Request Broker Architecture (COBRA), Abstract Syntax Notation Number 1(ASN.1) and Extensible Markup Language (XML).

Flat files contain machine-readable data that is typically encoded as printable characters. A flat file database usually contains a series of record, where each record is a sequence of fields. Much of it is stored in flat files as tab-delimited text, comma-separated values, or some similar format. The usage of flat file in protein database have a few problems such as accessing the data is limited to the single user, high probability of data damage, data control is very difficult and data are highly redundant. CORBA is a good solution for developing and deploying application in heterogeneous environments [4]. CORBA infrastructure was used for development at EMBL-EBI [8] and proof that the CORBA interface can address some of the limitation of the flat-files format. It provides means for accessing and distributing EMBL Data [5]. Disadvantages of CORBA are the construction of servers in CORBA is difficult and time consuming. Interface Definition language (IDL) places too many restriction on how to use the data and make it user hardly to use [7]. ASN.1 is also being used at NCBI to store data as well as send it and received it. The common known formats that are produced from NCBI database is GENBANK flat file. The disadvantages of ASN.1 are the values are in binary format; need the expertise to guide in order to understand the result [5]. XML is an emerging standard for storing biological data in a structured format [8]. XML is

are *seqfeature*, *bioentry*, *terms* and *relation*. B*iosqldb_mysql* script creates an instruction to create table, primary key, secondary key and determine the relation between each table. Figure 3 shows the entity relationship diagram for the *protein_info*.

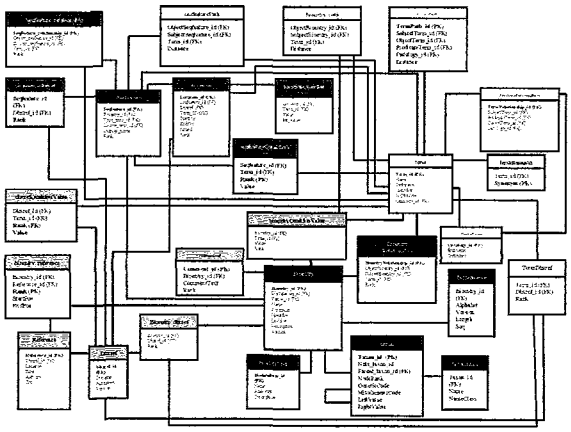

Fig. 3 Entity Relationship diagram for the *Protein_info*

C. Display Result based on User Query

We provide a searching tool using bioperl[9] and SQL. This tool is used to find/search the sequence and other information in the *protein_info* database based on sequence, id or species name.

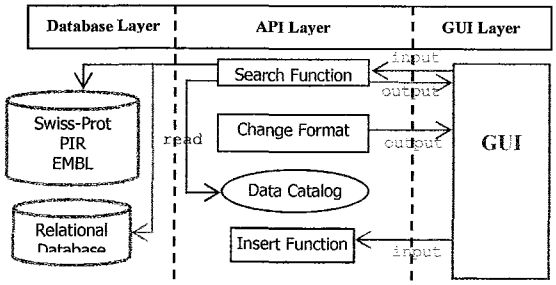

Fig. 4 System's architecture

From the figure 4, the user can access the system via web to manipulate and search the information from the relational database i.e *Protein_info* by using Search Function. We used CGI/Perl to access the Bioperlmodules. The Search Function, Insert function, Change format and Data Catalog are the Bioperl functions.

IV. Implementation

We implemented this system on a PC cluster machine called TIARA installed with Apache HTTP server. The tool is written using Bioperl modules and C/C++ programs. The graphical user interface uses HTML and JAVA script with CGI Perl to interface with the backend program. MySQL is used to store the protein data. The Web based system is available at http://tiara.cs.usm.my/Bioprotein.html. We have used Swiss-Prot, EMBL, PIR as sequence database.

As described at figure 2 in section III, this system will map protein flat files from the flat file databases into XML and changed into a relational database. The conversion of the information from different databases will be transformed into one standard format. We have developed one *rules file* to store all mapped information. The mapped information is written into a file with extension *.swissprot* for Swissprot database, *.Pir* for pir database and etc. The mapping file is created only once and can be used every time user wants to change the protein information into the database.

The *rule file* is created with the protein flat file. The *bean file* is also created to extract out the respective fields of information from different protein databases. Each of the protein databases has their own *bean file*. By using Doc type, XML *Out putter* is used to generate XML. Generated XML file with predefined DTD will be used to validate the XML. Data will be inserted to the relational database if the XML file is valid.

Once the protein flat files is converted into XML and validated, the data can be inserted into relational database i.e *protein_info* which has been constructed as mentioned in section III. In order to evaluate and compare the performance of the system, search module is provided. Each of the queries to *protein_info* using SQL forms in the Perl Language. Searching module is developed to search sequences and other information based on sequence, ID or species requested by the user. Example of protein sequence: MASVKSFILILSQVIECQPOS, example of ID protein: 108_LYCES and example of species is proteinTheileria parva.

V. Experimental Results

The performance of search is evaluated using various measurements such as reliability, accuracy and performance. Each of this measurement will be discussed in details. 100000 data protein from difference protein databases have been downloaded into the Tiara server. In this system, the final results are displayed through search result. The analysis will be focused on search module.

i. Reliability

Reliability of the information was evaluated by comparing the search results in relational database with the original flat file database. There are 23 unique common names in the database. The features used to evaluate are contents of information. The result is same as in flat file format and all the values mapped able to retrieved from relational database is same as in flat files.

ii. Accuracy

Accuracy evaluation was done to determine the ratio of retrieved data protein is relevant. Execution of this evaluation

