A Software Package for Managing and Evaluating DNA Sequence and Microsatellite Data

Cong Truong, Zhivko Duchev and Eildert Groeneveld

Department of Breeding and Genetic Resources
Institute of Farm Animal Genetics, Mariensee, FLI
Höltystr 10, 31535 Neustadt, Germany
cong.chi@fli.bund.de, zhivko.duchev@fli.bund.de, eildert.groeneveld@fli.bund.de

Abstract: We have surveyed three molecular biology labs in Germany and Vietnam to evaluate practical problems in the management of molecular genetics data. These labs are generating a large amount of heterogeneous genetic data. Without a long-term and uniform storage as well as the assistance of proper statistical analysis tools, the management and evaluation of experimental data has become difficult. Based on the formalized workflow and collected data elements, a common data model is created. A Web application is designed and developed to provide all essential features of a Laboratory Information Management System (LIMS). The system can be used to manage and evaluate DNA sequence and microsatellite data in both plant and animal breeding for all different species. Under Open Source GNU public license, our software package will be released via the virtualization technology.

1 Introduction

Recent technological advances in molecular biology and genetics [Vi02] have rapidly increased the output of scientific data. Molecular markers, for instance, SSRs (simple sequence repeats) and SNPs (single nucleotide polymorphisms) [RSM05] are extensively applied in genetic research. Therefore, the need for a long-term storage and effective management of massive volumes of genetic materials, samples and experimental results has become a major issue in molecular genetics labs. At present, many labs still use traditional methods (e.g. workbooks, spreadsheets) to handle their data while software providers are striving to advertise their software. Buying a Laboratories Information Management System (LIMS) such as Biotracker, Scierra Sequencing, Genetell and Identitrack, SQL*LIMS, Modul-Bio, Geneus is a challenge because most of commercial products in the field of molecular biology are expensive. In other words, this approach requires sufficient investment of funds, which is not always available for small labs [Vi07]. Developing in-house software is an alternative approach to meet custom requirements. But the information technology experts must be available in your lab. Besides, costs in time for designing, implementing and testing a software must be considered. Many free solutions were developed and discussed in Mogelis [Sw04] for management of microarray projects while dealing with data from projects about DNA sequencing and microsatellite genotyping is seldom. Here we want to find common tasks in the workflow in biology or molecular genetics labs thereby allowing the same software to be used across all of the labs without modification. Our project is an attempt to address data integration from disparate sources. In this project, we want to develop an open source information system to efficiently handle such large volumes of data.

2 Objectives and Methods

Our major objective is to contribute a free software package which can be used to collect, manage and evaluate molecular genetics data. The data model used in our information system has to be designed at the formalized level in order to cover essential needs for various labs. Specifically, we is trying to achieve the following requirements:

1. The software package is directed toward DNA sequencing and microsatellite genotyping projects in small molecular genetics labs.

2. Sample management is required as one of basic features to track all samples which are created and shared from different projects by different users.

3. The information system should capture all original data at earliest possible stage.

4. All different outputs which are created from different instruments or machines should be stored in one uniform format in order to be evaluated easily.

5. The software package supports the researchers to record and keep track of their data as well as experimental results at each step in their workflow.

6. Data entry, searching, analysing and reporting have to be implemented with a high degree of automation.

7. The information system must be a multi-user system which supports security and access control.

8. The software package can be deployed for several platforms such as Unix, Linux, Windows.

9. The data model must be generalized to cover common needs of different labs.

10. The information system can be developed from various open source packages and being also open source software released under GPL to allow everyone install, use, distribute, and modify without any software license cost.

11. English must be used as the default language for both application program and documentation.

3 Architecture and Implementation

As pointed out above, the information system will be an open source software package, therefore it must contain only open source codes. In fact, we have inherited most of the processing capacity in APIIS [Gr04], a framework for adaptable platform independent information systems, to develop our software. Hence, data modeling and application architecture are also driven from this framework.

Based on workflows in labs in Germany and Vietnam, data processing procedures at each step of an experiment were collected in order to create a formalized workflow for management of molecular genetics data presented in MolabWF [CDG08]. We built a common data model from the data framework in MolabWF. It is an integrated data model which can meet general requirements in terms of data management and retrievals. The data model is formed from three data components. First, *core data* are data entities required in APIIS and basic data components used in CryoWEB [CDG08] for managing national genebanks. Second, *workflow data* are data components used to keep track of data processing procedures at each step in the workflow. Third, *experiment results* are data components for recording data elements generated from experiments (e.g. organism information, sample data, gel images, protocol files, raw data, processed results).

Designed with a three-tier Client/Server architecture, our system provides basic features of a web-based application. The user interface tier has a web layout which is compatible with the W3C standard in order to work with many different web browsers (e.g. Firefox, Opera). However, the user can use a non W3C browser like IE as well. In the business logic tier, application programs written in Perl programming language are developed by using various Perl modules which are available freely on CPAN. Thus, the combination of CGI::Ajax and HTML::Templates gives us a proper solution to handle all dynamic forms in the same manner. In addition, Inline::Java package is used to integrate Jasper-Reports, an open source reporting library written in Java, into our application to compile and generate reports automatically. The templates of our reports can be designed and customized via iReport package. Besides, we use Prototype (an open source Javascript Framework) to control web layouts and dynamic interactions at the client side. PostgreSQL is used as the default backend in the database access tier.

As a result, the software package provides a friendly graphical user interface with a menu bar (Fig. 1) which allows users to interact with the system easily. The system also supports the workflow management allowing one process at one step is pipelined from a previous step. The system supports batch loading for inserting a large set of genetic materials or DNA samples. In addition, the data forms are optimized for the efficient data entry. Data capturing is done at each step in the workflow. These data can be updated or removed then by using the feature of data management. The report engine allows the user to compile and extract many kinds of different reports by project, samples, or time. Another feature of the system is that the user can also export his/her data to various formats called "input files" for statistical analysis software. This feature brings, indeed, practical benefits for research scientists because it reduces the time for data preparation and avoid human being errors. Moreover, the application supports administrative features which helps the administrator to be able to define data indexes or protocols, to update contacts' information, to change system variables and to set up the storage locations in his/her lab.

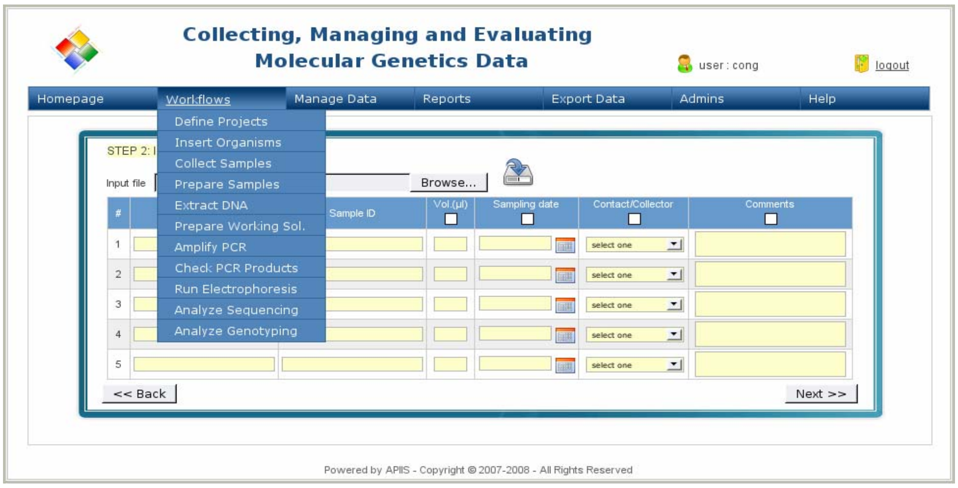

Figure 1: The graphical user interface for workflow management

4 Conclusions

We have developed an open source software package for managing and evaluating molecular genetics data. Our data model and application implementation are applicable for many small labs on the worldwide. Implemented as a Web application, the software package not only meet our requirements as stated above but also provide the ability to keep track of data via the workflow management. To cut down hardware costs as well as installation and configuration time, our software package will be deployed via the virtualization technology. We hope that many scientists working in the field of molecular genetics will use and evaluate this software package.

References

[Vi02] Vignal, A. et.al.: A review on SNP and other types of molecular markers and their use in animal genetics. Genet. Sel. Evol. 2002; 275-305.

[Gr04] Groeneveld, E.: An adaptable platform independent information system in animal agriculture: Framework and generic database structure. Livestock Production Science, 2004; vol. 87, pp. 1-12.

[Sw04] Swertz M. A. et.al.: Molecular genetics information system (molgenis): alternatives in developing local experimental genomics databases. Bioinformatics, 2004; vol. 20, no. 13, pp. 2075–2083.

[RSM05] Rudd, S.; Schoof, H.; Mayer, K: PlantMarkers: a database of predicted molecular markers from plants. Nucleic Acids Res, 2005; D628–D632.

[Vi07] Viksna, J., et.al: PASSIM – an open source software system for managing information in biomedical studies. BMC Bioinformatics. 2007; 8:52, 7 pp.

[CDG08] Cong, T. V. C.; Duchev, Z. I.; Groeneveld, E.: CryoWEB – A web application for managing national genebanks. DGfZ and GfT, 2008; pp. D7.

[CDG08]Cong, T. V. C.; Duchev, Z. I.; Groeneveld, E.: A formalized workflow for management of molecular genetics data. RIVF - IEEE, 2008; pp. 235-238.